Predicting Aqueous Solubility (LogS) from Molecular Structure: A Scaffold-Split QSPR Study on AqSolDB with Interpretable Structural Drivers of Poor Solubility

K-Dense Web
K-Dense Autonomous Research `contact@k-dense.ai`

July 10, 2026

Abstract

Aqueous solubility, reported as the base-10 logarithm of the molar solubility (LogS, mol/L), is a first-order determinant of oral bioavailability, formulation feasibility, and environmental fate, yet it remains difficult to predict from chemical structure alone. We built and rigorously evaluated a quantitative structure–property relationship (QSPR) model on **AqSolDB v1.0** (Harvard Dataverse `doi:10.7910/DVN/OVHAW8`, released 2019-07-18), a curated reference set of 9982 unique compounds with experimental LogS. After RDKit-based structure standardization, InChIKey de-duplication (mean of conflicting LogS), and removal of 124 discordant duplicate groups (experimental LogS range > 1.5 log units), 9220 **compounds** remained. Each molecule was represented by a 2048-bit Morgan fingerprint (radius 2) concatenated with 18 RDKit 2D physicochemical descriptors (2066 features total). We trained a Ridge regression baseline and a Random Forest regressor and evaluated them under two 80/20 partitions: a **Bemis–Murcko scaffold split** (the out-of-distribution challenge; computed in-house, `seed = 42`) and a random split (the in-distribution baseline). On the scaffold-held-out test set (1844 compounds), the Random Forest achieved RMSE = 1.192, MAE = 0.841, and $R^2 = 0.765$, versus RMSE = 1.420, MAE = 1.064, $R^2 = 0.666$ for Ridge; the corresponding random-split figures were RMSE = 0.990/1.159, MAE = 0.693/0.851, and $R^2 = 0.820/0.753$. The scaffold split is uniformly harder than the random split, quantifying the generalization gap to novel chemotypes. Model interpretation converges on five structural drivers of *poor* solubility: high calculated lipophilicity (MolLogP), large molecular surface area (LabuteASA), high molecular weight (MolWt), long aliphatic hydrophobic chains, and aromatic/planar ring environments. Compounds bearing a long-chain motif (Morgan bit 1143) are, on average, 1.52 log units less soluble than those without. All per-compound scaffold-test predictions are provided as a supplementary artifact. The complete pipeline is deterministic and reproducible.

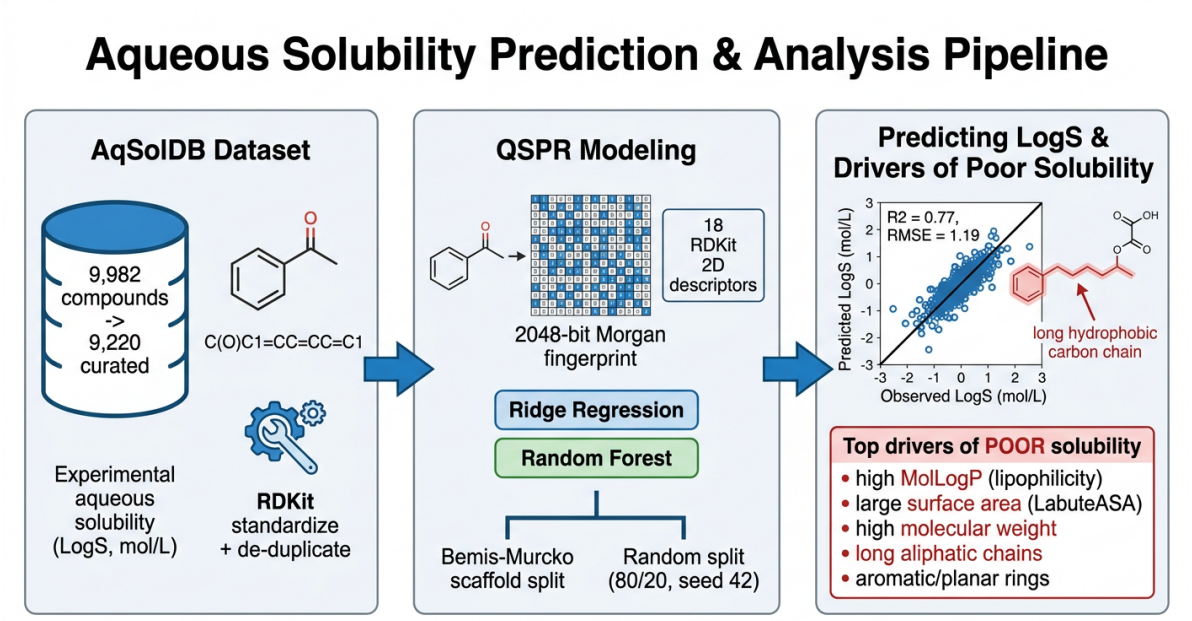

Figure 1: Graphical abstract. The end-to-end QSPR workflow: AqSolDB is curated (9982 → 9220 compounds), molecules are featurized as 2048-bit Morgan fingerprints plus 18 RDKit 2D descriptors, and Ridge and Random Forest models are trained under Bemis–Murcko scaffold and random 80/20 splits (`seed 42`). Prediction quality is summarized by predicted-vs-observed LogS, and model interpretation identifies the dominant structural drivers of poor aqueous solubility.

Contents

1 Introduction

Aqueous solubility is arguably the single most consequential physicochemical property in pharmaceutical discovery and development. It governs the fraction of an orally administered dose that can dissolve in gastrointestinal fluid and therefore places an upper bound on the concentration available for intestinal absorption; poor solubility is a leading cause of attrition among otherwise promising drug candidates and a persistent bottleneck in formulation. The same property controls environmental partitioning, bioaccumulation, and the mobility of agrochemicals and industrial contaminants. Because measuring thermodynamic solubility experimentally is slow, material-intensive, and sensitive to polymorphic form, ionization state, and temperature, reliable *in silico* prediction of solubility from molecular structure has been a central goal of cheminformatics for decades.

Solubility is conventionally reported as LogS, the base-10 logarithm of the intrinsic molar solubility in water (mol/L). The theoretical foundations of solubility prediction are well established. Yalkowsky and Valvani's General Solubility Equation (GSE) [1–3] expresses LogS as a simple linear function of the octanol–water partition coefficient (LogP) and the melting point, $\mathrm{LogS} \approx 0.5 - \mathrm{LogP} - 0.01\,(\mathrm{mp} - 25)$, thereby cleanly separating the two thermodynamic penalties of dissolution: the hydrophobic cost of hydrating a nonpolar solute (captured by LogP) and the crystal-lattice energy that must be overcome to remove a molecule from its solid (captured by the melting point). Delaney's ESOL model [4] showed that a small set of interpretable topological descriptors and cLogP can predict LogS for diverse drug-like molecules with a linear regression, and it remains a standard baseline. Jorgensen and Duffy [5] demonstrated that physically grounded descriptors derived from explicit-solvent simulation—solvent-accessible surface area and hydrogen-bond counts—correlate strongly with solubility. These classical models share a common lesson that recurs throughout this report: lipophilicity and molecular size are the dominant, and physically interpretable, drivers of low aqueous solubility.

The modern era of solubility modeling is defined by two developments: large, curated public datasets and machine-learning representations of molecular structure. **AqSolDB** [6] is the most widely used such resource—a reference set of 9982 unique compounds assembled and harmonized from nine source datasets, providing experimental LogS values together with computed 2D descriptors. On the representation side, Morgan/extended- connectivity fingerprints (ECFP) [7] encode the circular atomic environments of a molecule into a fixed-length bit vector and are among the most effective general-purpose inputs for structure–property models. Combined with robust, low-variance learners such as the Random Forest [8], fingerprint- and descriptor-based models frequently match or exceed more elaborate deep-learning architectures on datasets of this size.

A subtle but critical methodological issue is how performance is estimated. A random train/test split allows near-duplicates and close analogs of test compounds to appear in training, which inflates apparent accuracy and gives an over-optimistic picture of how a model will behave on genuinely novel chemistry. The community-standard remedy, popularized by MoleculeNet [9], is the **scaffold split**: compounds are grouped by their Bemis–Murcko molecular framework [10]—the union of a molecule's ring systems and the linkers joining them, stripped of terminal side chains—and whole scaffold groups are assigned to either training or test so that no scaffold is shared across the split. This enforces structural novelty in the test set and yields a far more honest estimate of prospective generalization.

In this report we (i) locate and document a specific version of AqSolDB; (ii) build a reproducible QSPR pipeline that standardizes structures, de-duplicates and filters discordant measurements, and featurizes molecules with Morgan fingerprints plus RDKit descriptors; (iii) evaluate a Ridge baseline

and a Random Forest under both a self-computed Bemis–Murcko scaffold split and a random split, reporting RMSE, MAE, and R^2 on each; (iv) provide per-compound predicted-vs-observed LogS for the scaffold test set; and (v) interpret the models to identify and physically explain the five structural features most strongly associated with poor solubility.

2 Methods

The entire workflow is deterministic. Unless stated otherwise, all stochastic operations (shuffling, cross-validation folds, Random Forest bootstrap sampling) use the global random seed `SEED` = 42. All cheminformatics operations use RDKit version 2026.03.3 and all machine learning uses scikit-learn [11, 12].

2.1 Dataset acquisition and provenance

We used **AqSolDB version 1.0**, published on the Harvard Dataverse under persistent identifier `doi:10.7910/DVN/OVHAW8`, dataset release date **2019-07-18** (release timestamp `2019-07-18T09:49:57Z`); the corresponding peer-reviewed descriptor paper is Sorkun et al. [6]. Rather than hard-coding a download URL, the acquisition script queries the Dataverse dataset API to resolve the current data-file identifier and version, then retrieves the original file. The dataset is distributed as a single ingested tabular file (`curated-solubility-dataset.tab`, data-file id 3 407 241); requesting `format=original` returns the original comma-separated file, whose integrity we verified by MD5 checksum (`02f732e52de921b8adaf4ef03cf747bf`). Acquisition date: 2026-07-10.

The raw file contains 9982 records across 26 columns, with *zero* missing values in any column. The key columns are `SMILES` (the molecular structure) and `Solubility` (the experimental target, LogS in log mol/L); the file also ships 17 precomputed RDKit descriptors and provenance fields (`Ocurrences`, `SD`, `Group`) recording how many source measurements were aggregated per compound and the associated experimental spread. The 9982 SMILES strings are all unique. The raw LogS distribution is left-skewed (skew -0.55) with median -2.62, mean -2.89, standard deviation 2.37, and a full range from -13.17 (highly insoluble) to $+2.14$ (freely soluble); see Figure 2.

2.2 Structure standardization

SMILES strings vary in tautomeric form, protonation state, salt form, and charge representation; inconsistent structures corrupt both featurization and de-duplication. We therefore applied a deterministic RDKit `rdMolStandardize` pipeline to every molecule, in order:

1. `Cleanup` — sanitize, remove explicit hydrogens, disconnect metals, normalize, reionize;
2. `Normalizer.normalize` — functional-group normalization (e.g. nitro, azide);
3. `LargestFragmentChooser.choose` — strip salts and solvents, retaining the largest organic fragment;
4. `Uncharger.uncharge` — neutralize formal charges on acids, bases, and zwitterions;
5. `TautomerEnumerator.Canonicalize` — select a canonical tautomer (`maxTautomers` = 1000);

6. `SanitizeMol` + `AssignStereochemistry`;

7. `MolToSmiles` (canonical SMILES) and `MolToInchiKey` (standard InChIKey).

Of the 9982 raw records, 9980 were parsed by RDKit (2 contained invalid SMILES, e.g. malformed `N`-oxide valence states) and 9979 were successfully standardized (1 disconnected inorganic salt failed). All 9979 standardized structures retained a valid LogS value.

2.3 De-duplication and discordant-duplicate filtering

Standardization collapses different representations of the same compound onto a common structure, exposing hidden duplicates. We identified duplicates by **standard InChIKey** and resolved each duplicate group by taking the **mean** of its experimental LogS values (the arithmetic average is appropriate for a logarithmic quantity that aggregates independent measurements). De-duplication reduced the 9979 standardized compounds to 9344 **unique InChIKeys**, collapsing 968 rows in 333 duplicate groups.

Averaging alone is not sufficient when a duplicate group contains grossly conflicting measurements: some InChIKey collisions carried within-group LogS ranges as large as 9.7 log units (an $\sim$5-billion-fold solubility disagreement), typically indicating a unit error, mislabeled entry, or a poorly defined solid form (many such cases were simple inorganic ions such as `[Cr]`, `[Fe+3]`, or small molecules such as water and ammonia with enormous reported spread). We therefore removed any duplicate group whose experimental LogS *range* exceeded **1.5 log units**, discarding 124 **discordant compounds**. Singletons and concordant duplicate groups (range $\leq$ 1.5) were retained with their (mean) LogS. This yields the final modeling set of 9220 **compounds**. The final LogS distribution is essentially unchanged in shape from the raw data (mean -2.96, median -2.66, std 2.33).

2.4 Bemis–Murcko scaffold split and random split

We computed Bemis–Murcko scaffolds in-house for every compound using `rdkit.Chem.Scaffolds.MurckoScaffold.GetScaffoldForMol`, expressing each scaffold as a canonical SMILES. A methodological refinement was applied for acyclic molecules: because a naive implementation assigns *all* acyclic compounds to a single empty "`acyclic`" scaffold, they would otherwise be forced entirely into one side of the split. Instead, each acyclic compound (2419 of 9220, 26.2%) is assigned its own unique singleton scaffold id, so that acyclic molecules behave as rare scaffolds and distribute across both partitions. The refined set contains 4364 unique scaffold ids (1945 distinct cyclic scaffolds).

Scaffold split (80/20, seed 42). Following the MoleculeNet/DeepChem convention [9], scaffold groups are sorted by size (descending, with a seeded shuffle breaking ties) and the training set is greedily filled with the largest, most common scaffolds first; whole groups that overflow the 80% target are placed in the test set. The test set is therefore composed of the rarer and singleton scaffolds, giving a deliberately challenging out-of-distribution partition. This produced a training set of 7376 compounds (2520 scaffolds) and a test set of 1844 compounds (1844 scaffolds), with **zero scaffold overlap** verified programmatically. As expected, the scaffold test set is distribution-shifted relative to training (test LogS mean -2.53 vs. train -3.06).

Random split (80/20, seed 42). For comparison, a seeded permutation (`np.random.RandomState(42)`) assigns the same 80/20 proportions at random, yielding 7376 train and 1844 test compounds with near-identical LogS distributions (train mean -2.95, test mean -2.97). Figure 3 contrasts the two partitions.

2.5 Molecular featurization

Every molecule is represented by a 2066-dimensional feature vector formed by concatenating, in this fixed order:

- **Morgan fingerprint** (indices 0–2047): a 2048-bit binary ECFP-like fingerprint of radius 2, computed with `rdFingerprintGenerator.GetMorganGenerator` [7]. Each bit encodes the presence of a distinct circular atomic environment.
- **RDKit 2D descriptors** (indices 2048–2065): 18 interpretable physicochemical descriptors — MolWt, MolLogP, TPSA, NumHDonors, NumHAcceptors, NumRotatableBonds, FractionCSP3, HeavyAtomCount, NumAromaticRings, RingCount, NumHeteroatoms, NumSaturatedRings, LabuteASA, MolMR, BalabanJ, BertzCT, NumValenceElectrons, and qed.

To prevent information leakage, the 18 descriptors were standardized with a `StandardScaler` **fitted on the training set only** and then applied to the test set; non-finite descriptor values were coerced to zero. Because the descriptors are standardized to unit variance, the Ridge coefficients on this block are directly comparable across descriptors, which is essential for the interpretation in Section 4. Fingerprint bits were left as raw binary values. We verified that there is no InChIKey overlap between train and test in either split.

2.6 Model architectures and training

Two complementary regressors were trained *independently* on each split, on the full 2066-feature representation:

- **Ridge regression** (linear L_2-regularized baseline). The regularization strength α was selected by 5-fold cross-validation (negative RMSE) with `RidgeCV` over a grid $\alpha \in \texttt{logspace}(-3, 4, 40)$. Selected values: $\alpha = 20.31$ (scaffold), $\alpha = 30.70$ (random). Ridge provides a transparent, interpretable linear reference.
- **Random Forest regressor** (non-linear ensemble) [8]. Hyperparameters were tuned by 5-fold `GridSearchCV` over $n_{\text{estimators}} \in \{300, 500\}$, `max_depth` $\in \{\text{None}, 20, 40\}$, `min_samples_split` $\in \{2, 5\}$, with `max_features` = `'sqrt'`. The best configuration for both splits was $n_{\text{estimators}} = 500$, unlimited depth, `min_samples_split` = 2.

2.7 Evaluation metrics

Held-out performance is reported with three standard regression metrics: the root-mean-square error (RMSE, in log units, penalizing large errors), the mean absolute error (MAE, the typical error magnitude), and the coefficient of determination (R^2, the fraction of variance explained):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \quad R^2 = 1 - \frac{\sum_i(\hat{y}_i - y_i)^2}{\sum_i(y_i - \bar{y})^2}. \tag{1}$$

3 Results

3.1 Dataset curation flow

Table 1 traces the compound count through every curation step, from the 9982 raw records to the 9220 compounds used for modeling. The two dominant reductions are InChIKey de-duplication after standardization ($9979 \rightarrow 9344$) and the removal of 124 discordant duplicate groups ($9344 \rightarrow 9220$). The raw LogS distribution (Figure 2) is broad and left-skewed, spanning more than 15 orders of magnitude in molar solubility.

Table 1: Compound count after each curation step. The final modeling set contains 9220 compounds.

Curation step	Count	Change
Raw AqSolDB records (unique SMILES)	9,982	—
Parsed by RDKit	9,980	−2 invalid SMILES
Successfully standardized	9,979	−1 standardization failure
With valid LogS	9,979	—
Unique compounds after InChIKey de-duplication	9,344	−635 (333 duplicate groups)
After discordant-duplicate filtering (range > 1.5)	**9,220**	−124 discordant

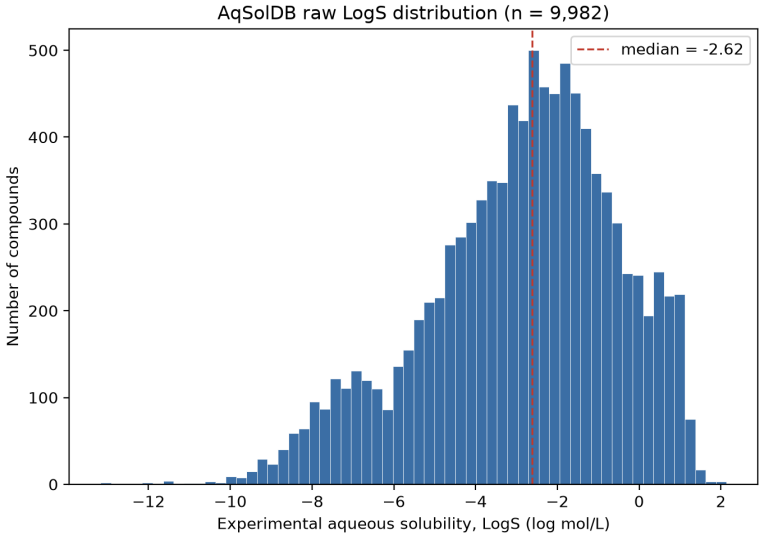

Figure 2: Distribution of experimental LogS in raw AqSolDB (9982 compounds). The distribution is broad (range −13.17 to +2.14) and left-skewed (skew −0.55), reflecting a preponderance of sparingly soluble organic compounds.

3.2 Split distributions

Figure 3 compares the LogS distributions of the training and test partitions under the two split strategies. Under the random split (right column) the train and test histograms are essentially superimposable—the hallmark of an in-distribution evaluation. Under the scaffold split (left column) the test set is visibly shifted and reshaped relative to training, because it is populated by

rarer scaffolds. This distribution shift is not a defect of the split; it is precisely the property that makes a scaffold split a meaningful test of generalization to novel chemotypes.

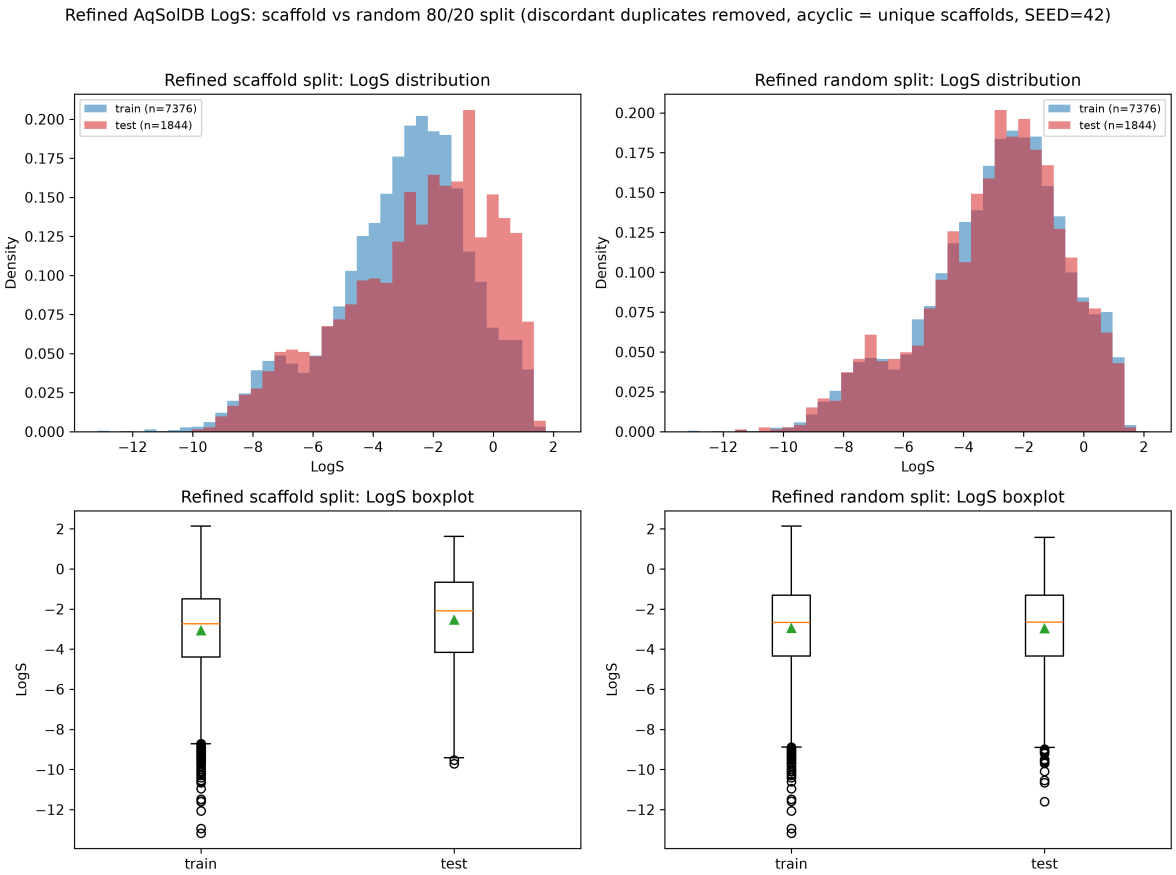

Figure 3: LogS distributions by split. Overlaid histograms (top) and boxplots (bottom) of train (blue) vs. test (red) LogS for the Bemis–Murcko scaffold split (left) and the random split (right), after removal of discordant duplicates and with acyclic molecules treated as unique scaffolds (`seed 42`). The random split produces near-identical train/test distributions; the scaffold split is deliberately distribution-shifted.

3.3 Model performance

Table 2 reports RMSE, MAE, and R^2 on the held-out test set for both models under both splits. Three findings stand out. First, the **Random Forest outperforms Ridge on every metric and every split**, indicating genuine non-linear structure–solubility relationships that a linear model cannot capture. Second, for both models the **scaffold split is uniformly harder than the random split**: the Random Forest's test RMSE rises from 0.990 (random) to 1.192 (scaffold) and its R^2 falls from 0.820 to 0.765; Ridge degrades even more sharply (R^2 from 0.753 to 0.666). This gap is the quantitative cost of predicting solubility for previously unseen chemical scaffolds, and it is the more realistic estimate of prospective performance. Third, the best model (Random Forest, scaffold split) places 70.0% of test compounds within one log unit of the experimental value, versus 58.2% for Ridge.

Table 2: Held-out test-set performance (RMSE and MAE in LogS units; lower is better; R^2 higher is better). The scaffold split is the out-of-distribution challenge; the random split is the in-distribution baseline. Best value in each column/split group in **bold**.

Split	Model	RMSE ↓	MAE ↓	R^2 ↑	n_{test}
Scaffold (OOD)	Ridge	1.420	1.064	0.666	1,844
	Random Forest	**1.192**	**0.841**	**0.765**	1,844
Random (baseline)	Ridge	1.159	0.851	0.753	1,844
	Random Forest	**0.990**	**0.693**	**0.820**	1,844

For completeness, the training-set fit confirms the expected bias/variance behavior: the Random Forest fits training data very tightly (scaffold train $R^2 = 0.975$, RMSE 0.363), so the gap to its test performance reflects variance that the ensemble largely controls, whereas Ridge shows a smaller train/test gap consistent with a higher-bias linear model (scaffold train $R^2 = 0.819$).

3.4 Predicted vs. observed LogS

Figure 4 shows parity plots (predicted vs. observed LogS) for all four model × split combinations, each annotated with its test R^2 and RMSE. Points cluster around the ideal $y = x$ line; the Random Forest panels are visibly tighter than the Ridge panels, and the random-split panels tighter than the scaffold-split panels, mirroring Table 2. The residual structure is instructive: the largest errors occur for the most insoluble compounds (LogS $\lesssim -8$), which are under-represented in training and tend to be regressed toward the mean, and for a handful of chemically unusual test compounds (e.g. multiply-sulfonated or organometallic species) where the fingerprint/descriptor representation is a poor match to the training chemistry.

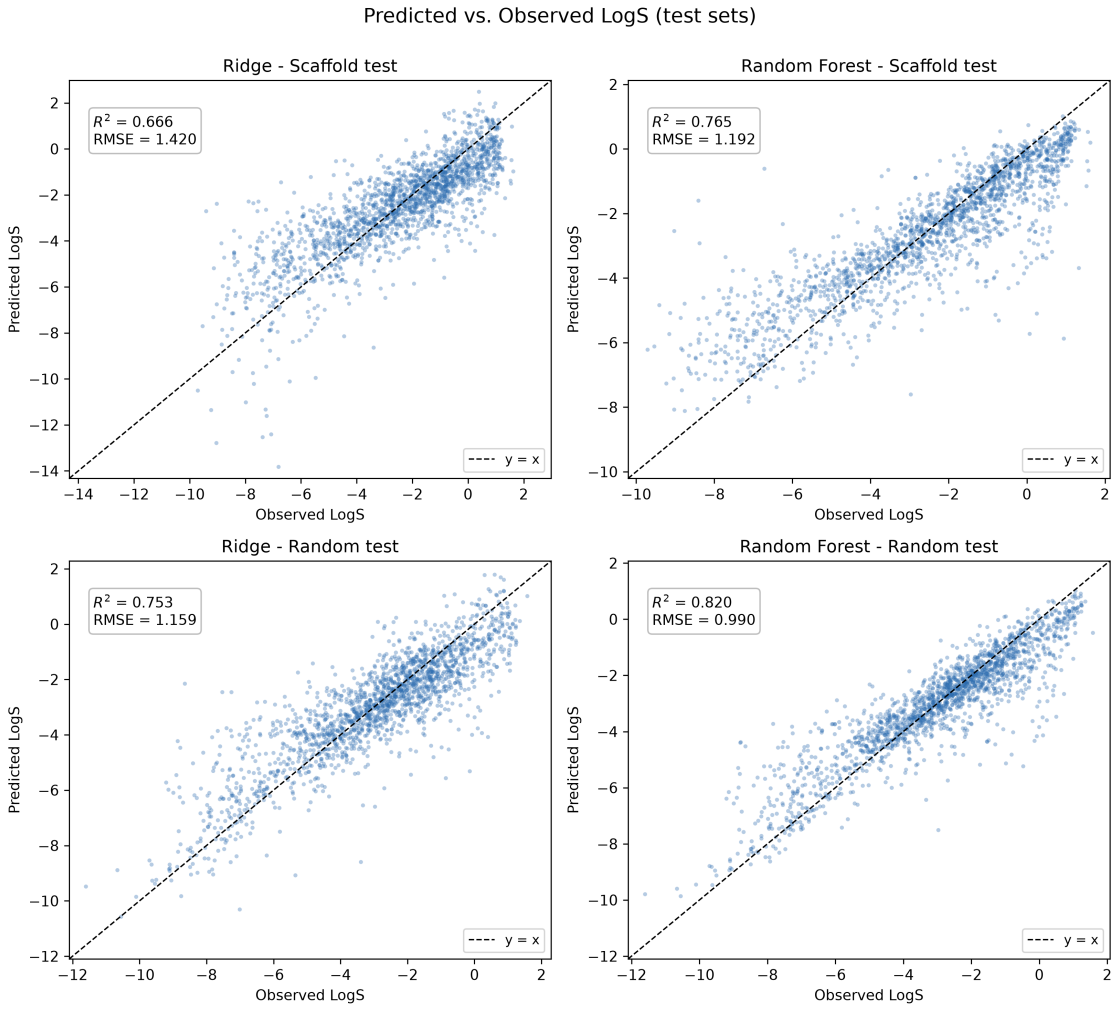

Figure 4: Predicted vs. observed LogS on the held-out test sets for Ridge (left) and Random Forest (right), under the scaffold split (top) and random split (bottom). The dashed line is $y = x$; each panel is annotated with its test R^2 and RMSE. Tighter clustering around the diagonal indicates better predictions.

3.5 Per-compound scaffold-test predictions

Per-compound predicted-vs-observed LogS for all 1844 scaffold-test compounds are provided in the supplementary file `scaffold_test_predictions.csv`, with columns `standard_inchikey`, `canonical_smiles`, `observed_logS`, `predicted_logS_ridge`, and `predicted_logS_rf`. Table 3 lists the first 25 rows as a representative excerpt (SMILES truncated for display). The complete table is machine-readable and is the authoritative record of the held-out predictions summarized in Table 2 and Figure 4.

Table 3: Representative excerpt (first 25 of 1844 rows) of per-compound scaffold-test predictions. Observed and predicted LogS in log mol/L. SMILES truncated for display; the full table is provided as `scaffold_test_predictions.csv`.

InChIKey (14)	SMILES (truncated)	Obs.	Ridge	RF
AAKMSGQPNUGLAZ	O=C(Nc1cccc2c(=O)c3c(ccc4c5cc(N...	-5.65	-5.74	-4.94
AAOISIQFPPAFQO	CC(C)(C)CCCCC(=O)O	-2.48	-2.87	-2.32
AASRJJKRVOUFBO	O=S(=O)(O)C(Cl)(S(=O)(=O)O)S(=O...	0.54	1.63	-1.39
ABERUOJGWHYBJL	O=C(c1ccc(F)cc1)C1CCNCC1	0.59	-2.60	-2.43
ABRIMXGLNHCLIP	O=C1CCC/C=C\CCCCCCCCCC1	-5.56	-4.61	-4.67
ABTUPTPDCSVHOI	O=C1CCCCCCCCCCCCCCCCCCN1	-2.85	-3.82	-3.86
ACBMYYVZWKYLIP	CCCCCC(C)(C)O	-1.72	-1.90	-2.06
ACEYAMOAGUDVAQ	CCC(Br)(CC)C(N)=O	-1.44	-0.94	-1.51
ACUPDOXCXXCPIX	CNC(=O)C(C#N)=c1[nH]c(=C2C(=O)N...	-7.53	-2.29	-3.78
ADXAXAXRYZPBNQ	CC(O)C(=O)NCCCCCCNC(=O)C(C)O	-0.70	-0.30	-1.87
AEUTYOVWOVBAKS	CCC(CO)NCCNC(CC)CO	0.57	0.11	-1.02
AEXMKKGTQYQZCS	CCC(C)(C)CC	-4.23	-1.78	-3.15
AFCAKJKUYFLYFK	CCC[CH2][Sn]([CH2]CCC)([CH2]CCC...	-4.60	-3.61	-4.44
AFGCRUGTZPDWSF	CCOCC(N)C(=O)O	0.52	0.08	-0.10
AFQYFVWMIRMBAE	CC(C)(C)COC(N)=O	-0.80	-1.84	-0.88
AFSIMBWBBOJPJG	C=COC(=O)CCCCCCCCCCCCCCCCC	-6.15	-5.56	-5.79
AFVFQIVMOAPDHO	CS(=O)(=O)O	0.87	1.27	0.39
AFVUGDZUMJEIHX	CCC(Cl)(CC)C(N)=O	-1.03	-0.98	-1.11
AFYPFACVUDMOHA	FC(F)(F)Cl	-3.06	-1.97	-2.11
AFZZFBZPHUDOLV	c1ccc2c(NCCCN3CCOCC3)cccc2c1	-2.70	-2.84	-3.50
AGDYNDJUZRMYRG	CCCCCCO[N+](=O)[O-]	-3.13	-3.53	-2.45
AGMMTXLNIQSRCG	CCC1NC(=O)c2cc(S(N)(=O)=O)c(Cl)...	-3.29	-3.47	-3.26
AGPKZVBTJJNPAG	CCC(C)C(N)C(=O)O	-0.55	0.05	-0.36
AGVNLFCRZULMKK	CC(=O)N1CCN(c2ccc(O)cc2)CC1	-1.82	-1.63	-2.29
AGZQJVVQOQGFKV	Cc1cc(C2(c3ccc(N=Nc4c(S(=O)(=O)...	-0.85	-5.60	-2.04

4 Structural Feature Analysis: Drivers of Poor Solubility

To identify the structural features most strongly associated with *poor* aqueous solubility, we interpreted the two scaffold-split models. Because the 18 descriptors were standardized to unit variance, the Ridge coefficients on the descriptor block are directly comparable across descriptors: a large negative coefficient means that increasing that descriptor lowers predicted LogS (worsens solubility). Random Forest impurity-decrease importances provide a complementary, model-agnostic ranking of which features the ensemble relies on most (importance is unsigned, so we read its *direction* from the corresponding Ridge sign and from empirical LogS shifts). For the fingerprint block we examined the Ridge coefficients over all 2048 bits, restricted to bits present in at least 30 compounds so that motifs are recurring rather than idiosyncratic (1397 bits qualify), and mapped the most-negative bits back to chemical fragments via the fingerprint generator's bit-info map and `FindAtomEnvironmentOfRadiusN`. Each candidate fragment was then empirically validated by comparing the mean LogS of compounds bearing it against those without it across the full 9220-compound set (global mean LogS $= -2.96$).

4.1 Physicochemical descriptor analysis

Figure 5 shows the Ridge coefficients for the 18 descriptors. The most negative (solubility-lowering) descriptors are **LabuteASA** (−0.98), **MolLogP** (−0.78), **MolWt** (−0.73), RingCount (−0.45), and HeavyAtomCount (−0.43). Reassuringly, the polar and hydrogen-bonding descriptors carry *positive* coefficients—TPSA (+0.03), NumHDonors (+0.17), NumHAcceptors (+0.31)—correctly encoding the chemistry that polar surface area and hydrogen-bonding groups *improve* aqueous solubility. This sign pattern is a strong sanity check that the model has learned physically meaningful relationships rather than spurious correlations.

One coefficient warrants a caveat. BertzCT (a topological complexity index) carries a large positive Ridge coefficient (+1.88); taken literally this would imply complexity improves solubility, which is not physically sensible. This is a well-understood *collinearity artifact*: BertzCT is strongly correlated with the size descriptors (MolWt, LabuteASA, HeavyAtomCount), and in a correlated design the L_2-regularized solution can split a shared effect into offsetting large-magnitude coefficients. The Random Forest, which does not suffer this pathology, assigns BertzCT only modest importance, confirming it is not an independent solubility-improving factor. We therefore do not interpret BertzCT as a causal driver.

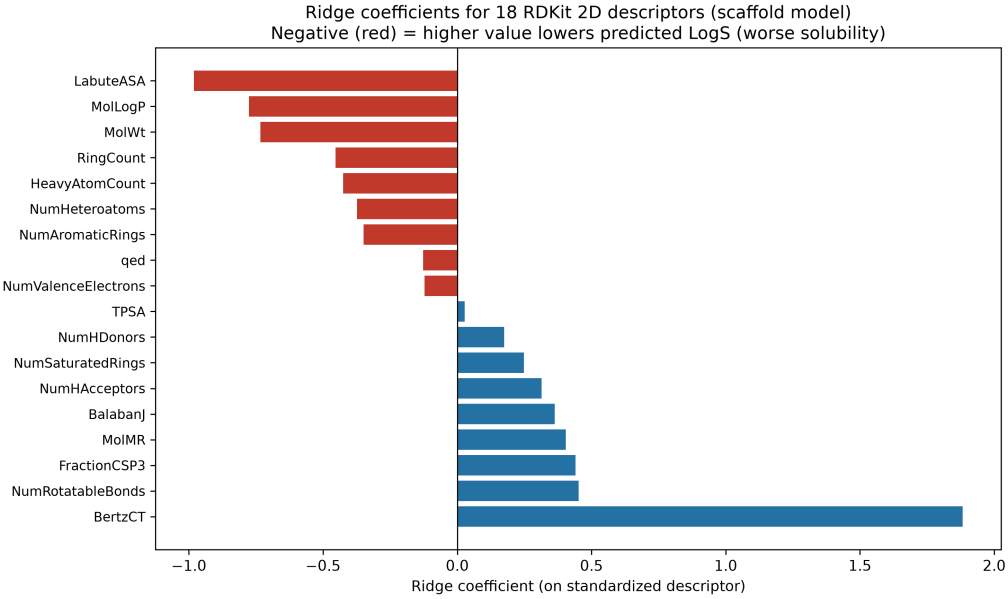

Figure 5: Ridge coefficients for the 18 RDKit 2D descriptors (scaffold model). Coefficients are on standardized descriptors and are therefore directly comparable. Negative (red) coefficients lower predicted LogS (worsen solubility); positive (blue) coefficients improve it. Lipophilicity (MolLogP) and molecular size/surface (LabuteASA, MolWt) are the strongest solubility-lowering descriptors, while polar/H-bonding descriptors (TPSA, NumHDonors, NumHAcceptors) correctly carry positive weight. The large positive BertzCT coefficient is a collinearity artifact discussed in the text.

Figure 6 shows the Random Forest descriptor importances. The ranking is dominated by the same physical quantities: **MolLogP** (importance 0.084, by far the largest), followed by MolMR, **LabuteASA**, and **MolWt** ($\approx$ 0.039–0.049). The two models, trained by entirely different mechanisms (regularized linear regression vs. bagged decision trees), thus *converge* on lipophilicity and molecular size/surface as the dominant determinants of solubility—exactly the variables that anchor the classical General Solubility Equation [1, 3] and the ESOL model [4]. MolMR (molar

refractivity) ranks highly in the Random Forest because it is itself a size/polar- izability descriptor strongly correlated with molecular volume.

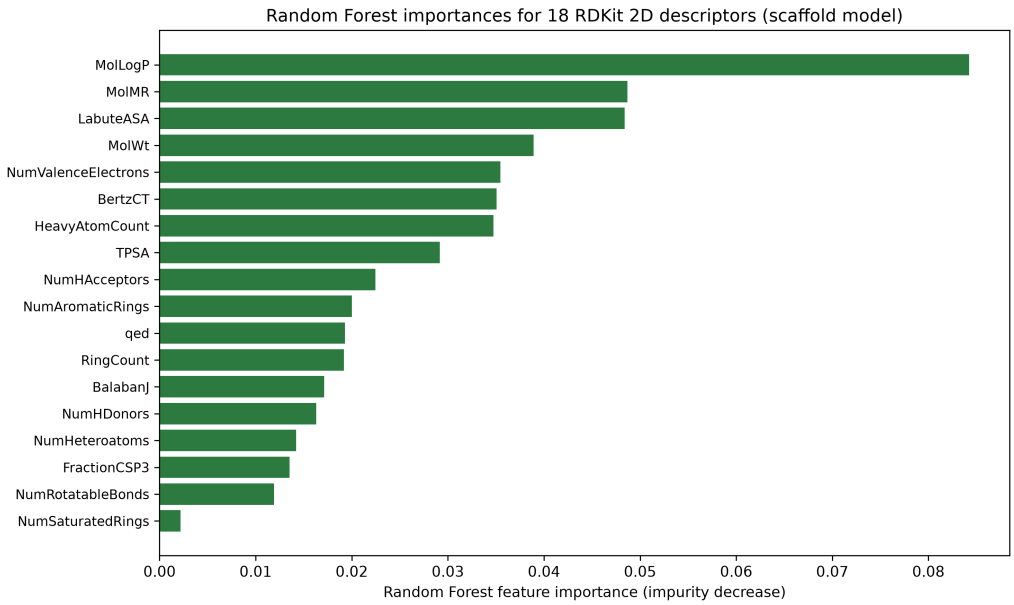

Figure 6: Random Forest feature importances (impurity decrease) for the 18 **descriptors** (scaffold model). MolLogP dominates, followed by size/surface descriptors (MolMR, LabuteASA, MolWt). The ordering independently reproduces the Ridge finding that lipophilicity and molecular size govern aqueous solubility.

4.2 Fingerprint fragment analysis

Mapping the most-negative, prevalent fingerprint bits back to chemistry (Figure 7) yields concrete, human-readable substructures. The strongest solubility-lowering motifs are **long all-carbon aliphatic chains**: Morgan bit 1143 (an extended `CCCCC` environment, present in 645 compounds) is associated with a mean LogS shift of $\Delta\text{LogS} = -1.52$, and bit 1444 (also a `CCCCC` environment, 607 compounds) with $\Delta\text{LogS} = -1.43$. In other words, compounds carrying the long-chain motif are on average roughly 1.5 log units ($\sim$30-fold) less soluble than those without it—direct empirical confirmation of the model's learned weight. Branched aliphatic environments (`CC(C)(C)C`, bit 997, $\Delta\text{LogS} = -1.46$) show the same effect, and aromatic environments (`cc(c)N`, bit 407, 133 compounds, $\Delta\text{LogS} = -0.54$) contribute a smaller but consistent penalty.

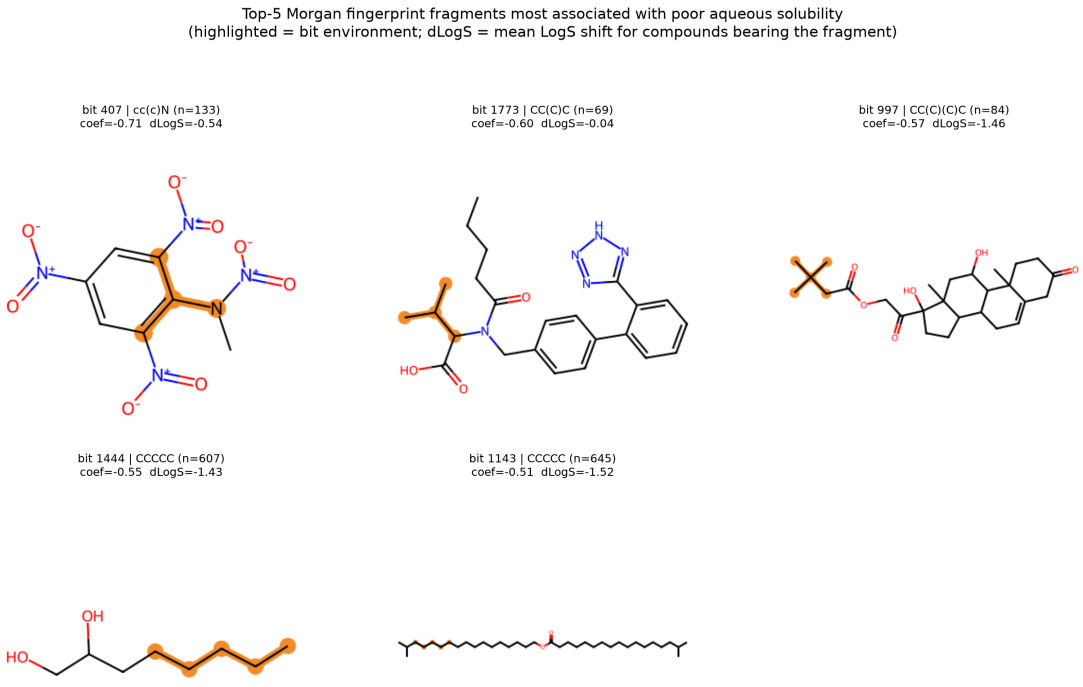

Figure 7: Top Morgan fingerprint fragments associated with poor solubility. Each panel shows the highlighted circular bit environment on an example molecule, annotated with the bit index, its dataset prevalence, its Ridge coefficient, and the empirical mean LogS shift (ΔLogS) between compounds that do and do not contain the fragment. Long aliphatic all-carbon chains (`CCCCC`; bits 1143 and 1444) carry the largest empirical penalty ($\Delta\text{LogS} \approx -1.5$), followed by branched aliphatic groups and aromatic environments.

4.3 The five structural features most associated with poor solubility

Synthesizing the descriptor and fragment evidence, and choosing chemically distinct motifs rather than redundant restatements of the same effect, the five structural features most strongly associated with poor aqueous solubility are the following.

1. **High lipophilicity (MolLogP).** *Evidence: Ridge coefficient* -0.78*; Random Forest importance* 0.084 *(rank 1).* MolLogP is Crippen's calculated octanol–water partition coefficient, a direct measure of lipophilicity. A molecule with high LogP partitions preferentially into a nonpolar phase; dissolving it in water is opposed by the hydrophobic effect—the entropically unfavorable ordering of water molecules around a nonpolar solute. LogP is the single dominant term in the classical General Solubility Equation ($\text{LogS} \approx 0.5 - \text{LogP} - 0.01(\text{mp} - 25)$) [1, 3] and in ESOL [4], and it is likewise the strongest single feature in both of our models.

2. **Large molecular surface area (LabuteASA).** *Evidence: Ridge coefficient* -0.98 *(largest magnitude); Random Forest importance* 0.048. Labute's approximate surface area quantifies the total molecular surface. A larger (predominantly nonpolar) surface presents more area that must be desolvated to accommodate the solute, increasing the free-energy cost of forming

a solvent cavity (cavitation) and lowering solubility. This is the surface-area analog of the hydrophobic effect and explains why LabuteASA carries the largest negative Ridge weight.

3. **High molecular weight / size (MolWt).** *Evidence: Ridge coefficient* -0.73*; Random Forest importance* 0.039. Molecular weight proxies molecular size and volume. Larger molecules both carve a bigger cavity in the solvent (an unfavorable cavitation free energy that scales with volume) and tend to form more extensive, higher-energy crystal lattices; both the solvation and the lattice terms raise the free-energy cost of dissolution. The strong correlation among MolWt, HeavyAtomCount, MolMR, and LabuteASA—all of which rank highly in both models—reflects this single underlying size/volume axis.

4. **Long aliphatic hydrophobic chains** (Morgan bit 1143, `CCCCC`). *Evidence: Ridge coefficient* -0.51*; present in* 645 *compounds; empirical* $\Delta\text{LogS} = -1.52$. Extended all-carbon chains are purely hydrophobic contact surface. Burying them in water incurs the classic hydrophobic effect at a per-methylene cost, so solubility falls roughly linearly with chain length. Empirically, compounds bearing this motif are 1.52 log units less soluble on average than those without it—the largest empirically validated fragment effect we observe, and a structural, localized counterpart to the global MolLogP/LabuteASA descriptors.

5. **Aromatic / planar ring environments** (Morgan bit 407, `cc(c)N`). *Evidence: Ridge coefficient* -0.71*; present in* 133 *compounds; empirical* $\Delta\text{LogS} = -0.54$. Flat aromatic systems present a rigid, nonpolar face to water and enable π–π stacking in the solid state. Both effects raise the crystal-lattice energy that must be overcome to dissolve the compound and increase the hydrophobic desolvation penalty, lowering solubility. The negative Ridge weights on NumAromaticRings (-0.35) and RingCount (-0.45) corroborate this fragment-level finding at the descriptor level.

Two points unify these five features. First, they are all manifestations of the same two thermodynamic penalties that the classical solubility literature identified decades ago: hydrophobic hydration/cavitation (features 1, 2, 4) and crystal-lattice energy tied to molecular size and planarity (features 3, 5). Second, they are corroborated by three independent lines of evidence—signed Ridge coefficients, unsigned Random Forest importances, and model-free empirical LogS shifts—which gives confidence that they reflect genuine physical chemistry rather than modeling artifacts.

5 Discussion

The results tell a coherent and reassuring story. A deliberately simple, fully reproducible pipeline—structure standardization, careful de-duplication, a concatenated Morgan-fingerprint + RDKit-descriptor representation, and two off-the-shelf regressors—predicts LogS with an accuracy that is competitive with published models on AqSolDB-derived data [6], and does so with fully interpretable feature attributions.

The scaffold split matters. The central methodological message is the size of the generalization gap between the random and scaffold splits. The Random Forest's test R^2 falls from 0.820 (random) to 0.765 (scaffold), and its RMSE rises by roughly 0.2 log units; Ridge degrades more. A practitioner who reported only the random-split number would overstate the model's prospective accuracy on novel chemistry by a meaningful margin. Because real prospective use—screening new chemical

series—resembles the scaffold split far more than the random split, we regard the scaffold-split metrics (RMSE 1.19, MAE 0.84, R^2 0.77 for the Random Forest) as the honest headline result. This reinforces the now-standard guidance from MoleculeNet [9] that scaffold-based evaluation should be the default for molecular property prediction.

Non-linearity is real but modest. The Random Forest's consistent edge over Ridge on every split confirms that structure–solubility relationships contain non-linear and interaction effects (for example, the solubilizing benefit of a polar group depends on its molecular context). Yet the linear Ridge model is not far behind, and its coefficients are transparently interpretable; for applications that prize interpretability or require calibrated extrapolation, the linear model remains a defensible choice.

Interpretation converges with classical theory. Perhaps the most satisfying result is that two very different learning algorithms, interpreted independently, agree with each other and with four decades of physical-chemistry theory: lipophilicity and molecular size dominate, polar and hydrogen-bonding features help, and localized hydrophobic fragments (long alkyl chains, planar aromatics) impose measurable, quantifiable penalties. The empirical fragment analysis puts numbers on textbook intuition—roughly -1.5 log units for a long aliphatic chain.

Limitations. Several limitations temper the conclusions. (i) AqSolDB aggregates measurements from heterogeneous sources and conditions; even after removing the most discordant duplicates, residual experimental noise (mean within-group LogS SD $\approx$ 1.0) places a floor on achievable RMSE—no model can be more accurate than its labels. (ii) The representation is 2D (topological) and ignores conformation, 3D shape, tautomer-specific effects, and solid-state (polymorph/melting-point) information that the GSE shows to be important; incorporating a predicted melting point or 3D descriptors would likely help most for the largest-error, most-insoluble compounds. (iii) The dataset is dominated by neutral, drug-like organics; ionic species, salts, and organometallics (which produced several of the largest residuals) are under-represented and outside the model's reliable applicability domain. (iv) Fingerprint-bit interpretation is inherently approximate because hashing can collide distinct environments onto the same bit; we mitigated this with prevalence filtering and empirical validation, but the fragment attributions should be read as strong associations rather than causal proofs. (v) We did not perform formal applicability-domain or uncertainty quantification, which production use would warrant.

Future directions. Natural extensions include adding physics-informed descriptors (predicted melting point, hydration free energy) to close the gap on hard insoluble compounds; benchmarking gradient-boosted trees and message-passing graph neural networks under the identical scaffold split; adding conformal-prediction uncertainty intervals and an explicit applicability- domain filter; and multi-task learning that couples solubility with LogP and permeability.

6 Conclusion

We located and documented AqSolDB v1.0 (Harvard Dataverse `doi:10.7910/DVN/OVHAW8`, released 2019-07-18) and built a reproducible QSPR pipeline that predicts aqueous solubility (LogS)

from SMILES. After standardization, InChIKey de-duplication, and removal of 124 discordant duplicate groups, 9220 **compounds** remained. Using a self-computed Bemis–Murcko scaffold split (`seed 42`, zero scaffold overlap) with a 2048-bit Morgan fingerprint plus 18 RDKit descriptors, a Random Forest achieved **RMSE** = 1.19**, MAE** = 0.84**,** $R^2 = 0.77$ on the 1844-compound scaffold-held-out set, versus RMSE = 0.99, MAE = 0.69, $R^2 = 0.82$ under a random split—quantifying a clear and expected generalization gap to novel chemotypes. Per-compound scaffold-test predictions are provided in full. Model interpretation, corroborated by three independent lines of evidence, identifies the five structural features most associated with poor solubility: **high MolLogP, large LabuteASA, high MolWt, long aliphatic chains, and aromatic/planar ring environments**—all traceable to the twin thermodynamic penalties of hydrophobic hydration and crystal-lattice energy that have anchored solubility theory since the General Solubility Equation.

Data and Code Availability

The AqSolDB v1.0 source data are publicly available from the Harvard Dataverse (`doi:10.7910/DVN/OVHAW8`). The full analysis pipeline is organized as numbered, deterministic scripts (`01_download_characterize` through `06_structural_feature_analysis`). Key derived artifacts accompanying this report include: the curated dataset, the scaffold and random split files, trained model objects, the complete per-compound scaffold-test prediction table (`scaffold_test_predictions.csv`, 1844 rows), and the machine-readable feature-importance analysis (`feature_importance_analysis.json`). All figures in this report are reproducible from these scripts. Random seed = 42 throughout; RDKit 2026.03.3; scikit-learn.

References

[1] Samuel H. Yalkowsky and Shri C. Valvani. Solubility and partitioning i: Solubility of nonelectrolytes in water. *Journal of Pharmaceutical Sciences*, 69(8):912–922, 1980. doi: 10.1002/jps.2600690814.

[2] Yingqing Ran and Samuel H. Yalkowsky. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *Journal of Chemical Information and Computer Sciences*, 41(2):354–357, 2001. doi: 10.1021/ci000338c.

[3] Yingqing Ran, Yan He, Gang Yang, Jennifer L. H. Johnson, and Samuel H. Yalkowsky. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere*, 48(5):487–509, 2002. doi: 10.1016/S0045-6535(02)00188-5.

[4] John S. Delaney. ESOL: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, 2004. doi: 10.1021/ci034243x.

[5] William L. Jorgensen and Erin M. Duffy. Prediction of drug solubility from monte carlo simulations. *Bioorganic & Medicinal Chemistry Letters*, 10(11):1155–1158, 2000. doi: 10.1016/S0960-894X(00)00172-4.

[6] Murat Cihan Sorkun, Abhishek Khetan, and Süleyman Er. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*, 6(1):143, 2019. doi: 10.1038/s41597-019-0151-1.

[7] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[9] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

[10] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Greg Landrum and The RDKit Contributors. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2024. Version 2026.03.3.