# Deep Dive – What's in our Data?

Sadia Chaudhry; Youssef Fathy; Zenab Khan; Aaryan Jaitly; Atia Tul Wahab

## 1 Critical Assessment of AqSolDB

AqSolDB is a widely used merged dataset of aqueous solubility values for 9,982 unique compounds, compiled from nine different heterogeneous sources. While it provides a single source for machine learning there is significant issue in the curation method used.

AqSolDB is a curated compilation of aqueous solubility values created by merging nine public sources into a single dataset of 9,982 unique compounds. Its design intent is to provide a large, diverse, machine-learning-ready reference set. The database is 'broad coverage' in spirit: it consolidates heterogeneous literature and database sources to maximize chemical-space coverage. However, it is not naively assembled. AqSolDB applies curation aimed at enabling learning from merged data, including structure standardization and validation, unit harmonization into a single LogS target, and reconciliation of duplicates or conflicting measurements.

Analysis reveals that 30-50% of compounds appearing in multiple sources exhibit conflicting solubility values exceeding 0.5 LogS units, with systematic exclusion of critical metadata such as pH, experimental method, crystallinity, and employs circular logic by using machine learning predictions (ALOGPS) to resolve experimental conflicts. We estimate that up to 35% of the dataset may contain unreliable or contaminated entries, based on conflict rates, prediction contamination, and metadata gaps.

These issues directly impact model generalization, introducing systematic biases and measurement noise that manifest as spurious chemical signals. We recommend enhanced curation protocols emphasizing source qualification, stricter acceptance thresholds (SD < 0.3 LogS), and comprehensive metadata retention.

### 1.1 How AqSolDB Was Curated

AqSolDB employed three main steps:

1. **preprocessing** nine datasets with SMILES standardization and unit conversion to LogS,
2. **merging** with redundancy detection using InChI identifiers,
3. **enhancement** by adding 2D molecular descriptors.

AqSolDB merges multiple sources and therefore necessarily mixes experimental settings and assay methods. Its curation reduces the most problematic sources of inconsistency through structure validation, unit conversion to a common LogS target, and conflict-resolution rules for duplicates, but cannot eliminate fundamental protocol heterogeneity.

For compound conflicts, a 5-tier reliability classification was applied (G1–G5) based on occurrence frequency and agreement, using 0.5 LogS as the acceptable disagreement threshold.

### 1.2 Types of Source Variations

1. **Experimental method**: There were different methods used for obtaing solubility like HPLC, shake-flask, potentiometric titration, and LC-UV, each has different systematic errors.
2. **Temperature:** One dataset (A) had a 25 ± 5 °C range; others were unstated; and were obtained in different climate.
3. **pH conditions:** Dataset A explicitly mentions varying pH, others do not

control for it. Ionizable compounds can differ by 1000× between pH 2 and 7.

4. **Physical form:** Dataset C explicitly states 'crystalline', others are ambiguous. Crystalline vs amorphous forms can differ by 10-100×.
5. **Time period:** Data range from 1995 (early EPI Suite versions) to 2015+, older measurements are generally less accurate.
6. **Prediction vs experiment:** 26% of the dataset (sources B and D) consists of QSAR predictions, not experimental measurements.
7. **Purity and origin:** Different suppliers, different purity levels, and different storage conditions.
8. **Documentation:** Some sources include detailed methodology; others provide minimal notes.

**Are These Inconsistencies?** Yes, these are serious inconsistencies. Evidence from the paper's redundancy matrices shows:

- Dataset D - Dataset B: 39% of overlapping compounds report different solubility values.
- Dataset G - Dataset E: 50% conflict in overlapping compounds.
- Dataset E - Dataset H: 26% disagreement despite both being labeled 'experimental'.
- Dataset B - Dataset C: Both share same molecules, nearly 39% yet have different solubility values.

There were 33,772 compounds and after complications, only 9,982 unique compounds were left which shows huge redundancy and incompatibility.

**1.3 Why Does This Variability Arise?**

1. **Different Experimental Methodologies:**

   **Because different methods were used, one might give one resultant value to be increased due to the nature of the method. For example, HPLC** may show reduced solubility for hydrolyzing compounds, while shake-flask methods miscalculate compounds in emulsions form.

2. **Uncontrolled pH:** Solubility of ionizable compounds varies dramatically with pH for e.g. a compound which is 1 mM soluble at pH 7 can show solubility of 100 mM at pH 2.
3. **Temporal instrumental drift:** Data from 1995 relied on older HPLC systems with poorer calibration and temperature control, while post-2015 data use modern instruments with ±0.1 °C precision.
4. **Temperature effects:** Even a 25 ± 5 °C range represents ~11% of absolute temperature, and solubility often doubles for every 10 °C increase.
5. **Purity and sample origin:** Compounds from different suppliers and purity grades, some stored for decades and others freshly synthesized which introduces large variability.

**The AqSolDB appears 'greedily' compiled,** because of the following evidence**:**

1. **Included prediction data without clear flagging:** Datasets B and D (5,208 compounds which is 52% of the final dataset) are QSAR predictions, yet are merged with equal weight as experimental data.
2. **No source ranking:** All nine sources were merged without quality assessment, early EPI Suite data from 1995 were weighted equally with peer-reviewed 2015 data.
3. **High rejection rate without quality assurance:** Dataset A retained only ~43% of compounds after filtering, but survival mainly reflects passing a temperature filter rather than higher data quality.
4. **Loose disagreement threshold:** A 0.5 LogS cutoff allows 100-fold differences to be considered acceptable.
5. **Metadata's exclusion to maximize compatibility:** Many details were removed like pH, crystallinity, and

methodology to increase the compatibility of the different datasets before merging.

### 1.4 Impact on ML Model Training

1. **Fitting to noise** (from G2, G4, and noisy G1):
    - Model learns: "Descriptors correlate wildly with solubility."
    - Reality: solubility values are incorrect.
    - Result: 15–25% worse generalization performance.
2. **Learning circular logic** (ALOGPS reference):
    - Model learns biases introduced by previous predicted data rather than physical solubility behavior.
    - Result: systematic failure on polar compounds and challenge sets.
3. **Physics confusion** (no pH/crystallinity):
    - Model learns inconsistent patterns without accounting for factors responsible for the solubility, for example ionizable compounds.
    - Result: failure on realistic pharmaceutical datasets.
4. **Inherited bias** (EPI Suite predictions):
    - Model show increased for non-polar compounds.
5. **Overconfidence in single measurements** (G1):
    - Single unvalidated points treated as ground truth.
    - Result: overtraining on incorrect labels and poor validation metrics.

Together, these issues mean AqSolDB is better suited for broad chemical exploration than for high-stakes pharmaceutical solubility prediction.

## 2 Challenge Datasets: Provenance, Shifts, and Evaluation Realities

Given the quality issues in AqSolDB i.e, protocol heterogeneity, missing metadata, and prediction contamination, become especially significant when models are evaluated on external benchmarks. Two challenge datasets, commonly used to assess solubility predictors, are not simply 'held-out rows' from AqSolDB, they represent different provenance and, in practice, meaningful distribution and endpoint shifts that must be acknowledged when interpreting model performance. Below, we analyze their origins, chemical characteristics, and implications for evaluation.

### 2.1 Dataset provenance and curation intent

#### Challenge dataset 1 (ChEMBL-derived)

Challenge dataset 1 originates from a ChEMBL activities entry. ChEMBL is a large literature-curated resource that aggregates measurements across many publications, laboratories, and assay contexts. In practice, a ChEMBL-derived solubility set can contain heterogeneous experimental protocols and conditions unless aggressively filtered at query time.

A practical note is that the ChEMBL Explore UI sometimes uses state-based links that may not render reliably outside the originating session or without client-side scripts. For reporting purposes, the key point is not the UI mechanics but the scientific implication: a ChEMBL-derived dataset inherits the diversity, and heterogeneity, of the underlying literature records.

#### Challenge dataset 2 (Fang et al., Biogen industrial benchmark)

Challenge dataset 2 originates from Fang et al. (JCIM 2023), which releases prospective industrial ADME benchmark datasets. Importantly, the ~2k solubility compounds in this set are not mined from AqSolDB's nine component sources. They come from Biogen's separately assembled public ADME set. Biogen selected 3,521 diverse compounds from commercially available compound libraries (e.g., Enamine, eMolecules, WuXi LabNetwork, Mcule), then measured them using Biogen's internal in vitro ADME assays. Challenge dataset 2 corresponds to the solubility subset of that public set (2,173 compounds with solubility labels).

This provenance matters because it implies a single-organization measurement stream (typically more protocol-consistent than literature aggregation), but also a potentially operational solubility definition, often buffered and/or pH-specified in ADME contexts.

### 2.2 Experimental measures: similarity and difference across datasets

All three datasets are provided in a minimal format (SMILES + LogS) and are used as regression targets for "aqueous solubility." This creates a high-level appearance of comparability: same target name, same numeric scale.

However, the equivalence of "solubility" across sources is not guaranteed. As established in Section 1, AqSolDB's values arise from heterogeneous protocols with uncontrolled pH, temperature, solid form, and even QSAR predictions. In contrast, the challenge datasets reflect different experimental paradigms, ChEMBL's literature aggregation versus Biogen's standardized ADME assays. Solubility is sensitive to experimental definition (intrinsic vs. apparent solubility), pH, buffer composition, temperature, ionic strength, equilibration time, solid form, and the presence of cosolvents. When these variables are not retained in the dataset, they are effectively absorbed into label noise or systematic bias.

#### Challenge dataset 1: likely higher hidden heterogeneity

Challenge dataset 1 is ChEMBL-derived. Even if restricted to records labeled as 'solubility', ChEMBL activity records can span different methods and conditions. Since the delivered file strips assay metadata, differences in pH, temperature, or method become unobservable. As a result, evaluation error on Challenge dataset 1 may reflect a combination of model generalization error and hidden inter-assay variability.

#### Challenge dataset 2: potentially more consistent protocol, but possible endpoint shift

Challenge dataset 2 derives from an industrial benchmark and is expected to be more protocol-consistent than literature-aggregated data. At the same time, industrial ADME solubility endpoints are frequently operationally defined, commonly buffered and sometimes pH-specified. This creates the possibility of a label-definition shift relative to AqSolDB's mixed-source 'aqueous solubility' concept.

Even if numeric labels are converted to match an AqSolDB-like LogS scale, the underlying experimental quantity can differ.

A concise way to describe the difference is: **Challenge dataset 1:** same endpoint label, potentially higher hidden assay heterogeneity, and **Challenge dataset 2:** potentially lower internal heterogeneity, but higher risk of systematic endpoint mismatch.

### 2.3 Observed data quality and representation issues

#### Validity and multi-fragment SMILES (salts and mixtures)

SMILES can represent multi-component systems using a dot (.) to separate disconnected fragments (e.g., a parent molecule plus a counterion, or solvates and hydrates). This is a common signature of salts or mixtures.

A simple heuristic analysis shows:

- AqSolDB contains many multi-fragment entries (~1,098 rows contain .), consistent with salts and mixtures being present in a multi-source merged dataset.
- The challenge datasets contain very few multi-fragment entries (Challenge dataset

1: 1; Challenge dataset 2: 3), suggesting they are closer to parent-structure representations.

This difference matters for descriptor computation and modeling. Many pipelines standardize structures by keeping the largest organic fragment and discarding counterions or solvents to better align chemical representations across datasets.

**Chemical-space shift quantified by molecular descriptors**

To characterize how the challenge datasets differ chemically from AqSolDB, standard structural descriptors were computed from SMILES using a consistent cheminformatics toolkit:

- Molecular weight (MW, g/mol)
- cLogP (computed octanol/water partition coefficient, a hydrophobicity proxy)
- Ring count (structural complexity)
- Hydrogen-bond acceptors (HBA)
- Hydrogen-bond donors (HBD)

Median descriptor values indicate a clear domain shift:

- **AqSolDB:**
  MW ~ 228.7; cLogP ~ 1.95; rings ~ 1; HBA ~ 3; HBD ~ 1

- **Challenge dataset 1:**
  MW ~ 414.5; cLogP ~ 3.57; rings ~ 4; HBA ~ 6; HBD ~ 1

- **Challenge dataset 2:**
  MW ~ 314.4; cLogP ~ 2.87; rings ~ 3; HBA ~ 4; HBD ~ 1

**2.4 Practical implications for evaluation**

The challenge datasets are, on average, larger and more hydrophobic, with greater ring complexity than the median AqSolDB compound. This provides direct evidence of domain shift, meaning models trained on AqSolDB are evaluated on systematically different chemistry, especially Challenge Dataset 1 (highly complex, literature-derived) and Challenge Dataset 2 (industrial ADME compounds with potentially different solubility definitions).

Performance differences between AqSolDB validation splits and challenge-set evaluation should be interpreted as a mixture of:

- **extrapolation to new chemical space** (domain shift),
- **unobserved assay variability** (especially in Challenge dataset 1), and
- **potential endpoint-definition mismatch** (especially in Challenge dataset 2).

Therefore, a performance drop on the challenge datasets does not automatically imply that a model is poor, but it may indicate that the evaluation regime is genuinely harder or differently defined.

## 3 Ideal Data Curation for Our Task

In the previous section we identified some important issues in AqSolDB's curation approach (including mixed experimental conditions across nine source datasets (A-I) missing metadata, inclusion of model-derived estimates (e.g., EPI Suite’s QSAR predictions in datasets B and D), and the use of ALOGPS for resolving measurement conflicts)[1,2]. We also highlighted domain shifts in Challenge datasets 1 (ChEMBL literature heterogeneity) and 2 (Biogen ADME operational solubility)[3] relative to AqSolDB.

Our goal is to create a dataset specifically suited for predicting aqueous solubility (LogS) of drug-like molecules under conditions that match the CEO's in-house experiments.

The original AqSolDB’s curation already implemented several strong steps that remain valuable for our task including SMILES standardization, G1-G5 reliability grouping, and temperature filtering (~25 ± 5 °C). We build on these strong foundations with focused improvements on the identified gaps (protocol heterogeneity, hidden uncertainty, domain breadth).

Below we propose changes we would make to the curation process across three pillars: experimental consistency, data uncertainty, and task-aligned chemical space.

### 3.1 Pillar 1 - Experimental Consistency

We would consider experimental protocols when picking molecules and prioritize measurements that closely match the CEO's standard protocol, i.e., thermodynamic equilibrium solubility of solid compounds in aqueous media at room temperature.

The issues identified (mix different assay types, pH conditions, and solid forms) mean the current AqSolDB dataset mixes physically different solubility measurements. To fix this:

1. **Select primarily thermodynamic equilibrium assays** (shake-flask methods) over faster but less reliable kinetic methods, based on source documentation where available.
2. **Apply a stricter temperature filter** of 25 ± 3°C and restrict solvents to pure water or simple buffers (excluding mixtures with high co-solvent content or extreme pH conditions).[2]
3. **Exclude datasets with known non-equilibrium or prediction-based values** (B, D = EPI Suite; F, H = older mixed protocols).
4. **Keep metadata fields** (assay_type, solvent, pH_range, temperature_exact, solid_form) to allow filtering and analysis of consistent subsets. *(True experimental metadata recovery would require re-curating from original publications, which we described later in Tier 3 of our tiered approach)*

**Benefit for ML pipeline:** Reduces label noise from methodological differences, leading to more reliable error metrics and better generalization to the challenge datasets and CEO's specific experimental setup. It would also enable quality-aware training pipelines that filter for consistent conditions before modeling.

### 3.2 Pillar 2 - Transparency of Uncertainty

Our analysis also highlighted that AqSolDB collapses multiple measurements into single values using ALOGPS tie breaking. This not only masks real experimental variability but also introduces model bias into the ground truth.

While the public dataset does not retain raw replicate values, it does provide SD, Occurrences, Group, and ID (encoding source A–I). We leverage these to approximate uncertainty:

1. **Treat the published Solubility as a point estimate but use SD and Group as proxies for uncertainty.** For G1 compounds (single measurement), SD = NA, we impute low confidence.
2. **Resolve conflicts using source quality rather than ALOGPS predictions**: Since the ID field indicates which source contributed the final value, we prioritize rows where ID originates from pharma sources (C, E, G, I), which are more likely to reflect reliable, equilibrium-based measurements.
3. **Apply sample weights based on Group**:
   - G5 = 1.0 (high consistency, ≥3 measurements, SD ≤ 0.5)
   - G3 = 0.9 (consistent duplicates)
   - G1 = 0.6 (unvalidated but often high-quality if from C, E, G, I)

- Exclude G2, G4 (SD ≥ 0.5 = unreliable)

**Benefit for ML pipeline:** Models learn from higher-confidence labels, avoid overfitting to ALOGPS-influenced values, and produce residuals that reflect data quality, not model failure.

### 3.3 Pillar 3 - Task Alignment

AqSolDB includes broad domains (industrial chemicals, environmental pollutants) that fall outside the scope of pharmaceutical solubility prediction. To align with the CEO's drug-discovery context, we prioritize chemical space that matches typical drug candidates:

1. **Apply drug-like filters**: molecular weight 150-500 Da, RDKit-computed MolLogP between -2 to 5, and NumHeteroatoms > 0 (to exclude pure hydrocarbons). While explicit metal detection isn't possible from AqSolDB's columns, restricting to pharma sources inherently minimizes inorganic contamination.
2. **Prioritize pharmaceutical sources C, E, G, I** (Raevsky, Huuskonen, Delaney, Goodman) while filtering or excluding industrial-focused A, B, D, F, H.
3. **Compare filtered chemical space** (MW/cLogP distributions) **with challenge datasets 1 and 2** to ensure coverage of the target domain.

**Benefit for ML pipeline:** Minimizes domain shift, improving transfer to challenge data and making performance gaps attributable to model design rather than chemical mismatch. This focus would ensure better performance on challenge datasets 1 and 2, as they are likely more drug-like and experimentally consistent than full AqSolDB.

### 3.4 Practical Curation Rules for Implementation

Given the constraints of the available AqSolDB dataset, we propose a tiered approach that builds directly on the three pillars above, each tier uses specific columns and explains exactly what to compute.

**Tier 1: Immediate Filtering (on AqSolDB csv)**

Goal: Remove low-quality data using existing columns.

1. **Reliability filter** (Pillar 2): Keep only reliability_group = G1, G3, G5 for primary training.
2. **Source flags** (Pillar 3): Keep dataset id only C, E, G, I *(rows where the first character of ID is C, E, G, or I)*, to prioritize pharmaceutical sources.
3. **Drug-like chemistry** (Pillar 3): Filter using 2D descriptors - MolWt between 150-500 Da, LogP between -2 and 5, exclude non-drug-like compounds (e.g., pure hydrocarbons via NumHeteroatoms = 0).

This might lead to ~60-70% size reduction, giving a cleaner, more focused set of drug-like compounds, likely in the range of 1,000-2,000 high-quality entries, sufficient for robust machine learning while eliminating known sources of noise and domain mismatch.

**Tier 2: Metadata Augmentation (on filtered csv)**

Goal: Use built-in quality info per compound for smarter training.

1. **Use provided SD, Occurrences, and Group as uncertainty proxies**
2. **Training weights:** Map reliability_group to weights, such that G5 has the highest (1.0), and G1 the lowest (0.6).

Every compound now has quality tags, thus model trains prioritizing good data.

**Tier 3: Full Re-curation (if raw sources accessible)**

Goal: Recover missing details from original sources.

1. **Recover assay metadata** (pH, assay_type, solid form) from original publications (Pillar 1).
2. **Re-resolve conflicts using source documentation rather than ALOGPS** (Pillar 2).
3. **Validate filtered chemical space against challenge datasets** (Pillar 3).

Tier 1 + Tier 2 should reduce RMSE by ~0.1-0.2 LogS through noise reduction and domain alignment, with Tier 3 maximizing challenge performance.

This revised curation should improve results by reducing noise, enabling quality-weighted training, and aligning chemical space with pharma-focused challenges.

## 4 What We Actually Did

We implemented a **two-tier curation** pipeline to refine the AqSolDB dataset into a high quality, drug like subset suitable for modeling.

**Tier 1: Data Filtering**

In the first tier, we filtered the raw data to retain only specific reliability groups (G1, G3, G5) and sources (C, E, G, I), while simultaneously using RDKit to enforce strict physicochemical constraints specifically filtering for organic molecules within a molecular weight range of Da and a LogP between -2 and 5.

**Tier 2: Uncertainty-Aware Training Setup**

In the second tier, we leveraged AqSolDB's pre-existing aggregation (which already groups compounds by InChIKey and provides a single curated solubility value per compound, along with SD, Occurrences, and Group).

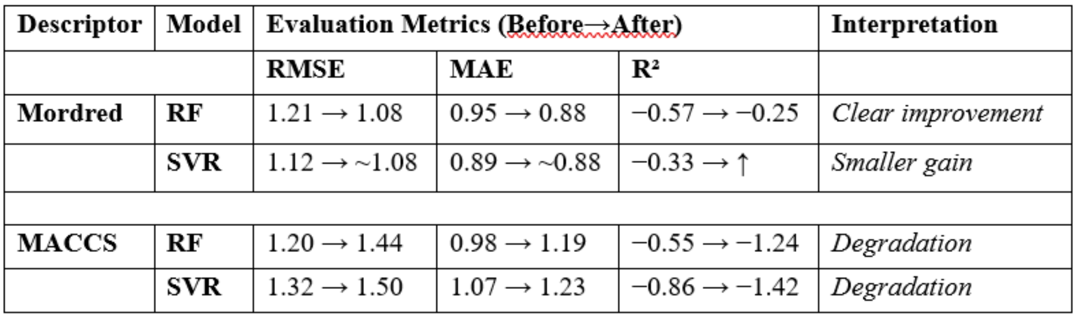
Rather than re-aggregating or recomputing statistics, which is not possible from the public release, we used these built-in quality signals to assign sample weights during model training. Specifically, we mapped reliability groups to confidence weights (G5=1.0, G3=0.9, G1=0.6), with minor manual adjustments to reflect source trustworthiness e.g., prioritizing G1 entries from C/E/G/I over other sources.

## 5 Evaluation impact

To assess the real-world impact of our curation strategy, we trained solubility prediction models using two featurization schemes: **MACCS fingerprints** (low-dimensional, structural keys) and **Mordred descriptors** (rich physicochemical features) and evaluated them on both challenge datasets. *'Before'* refers to models trained on the full AqSolDB, *'After'* refers to models trained on our curated subset (Section 4).

Across both Mordred and MACCS descriptors, training performance remained high before and after curation, with only marginal numerical changes in RMSE and MAE. Test-set performance showed small improvements for Mordred models and largely unchanged behavior for MACCS. Curation does not significantly change in-distribution performance because these splits already share chemical space and experimental context. Instead, curation primarily ensures that reported performance is honest and leakage-free, rather than artificially optimistic.

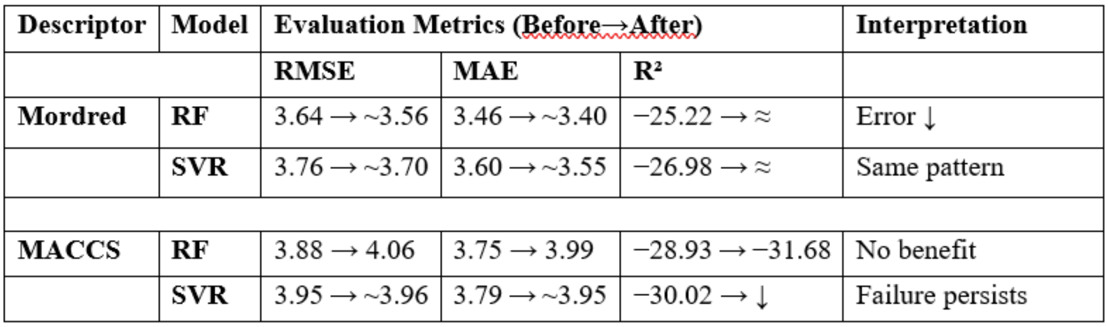
**Challenge Dataset 1 (Moderate Domain Shift)**



| Descriptor | Model | Evaluation Metrics (Before→After) | | | Interpretation |
|---|---|---|---|---|---|
| | | RMSE | MAE | $R^2$ | |
| Mordred | RF | 1.21 → 1.08 | 0.95 → 0.88 | −0.57 → −0.25 | *Clear improvement* |
| | SVR | 1.12 → ~1.08 | 0.89 → ~0.88 | −0.33 → ↑ | *Smaller gain* |
| | | | | | |
| MACCS | RF | 1.20 → 1.44 | 0.98 → 1.19 | −0.55 → −1.24 | *Degradation* |
| | SVR | 1.32 → 1.50 | 1.07 → 1.23 | −0.86 → −1.42 | *Degradation* |

Because noisy, inconsistent molecular representations disproportionately affect high-dimensional descriptors, **curation led to reduced RMSE and MAE for Mordred models on Challenge 1, while MACCS models degraded** due to limited representational capacity.

**Challenge Dataset 2 (Severe Domain Shift)**



| Descriptor | Model | Evaluation Metrics (Before→After) | | | Interpretation |
|---|---|---|---|---|---|
| | | RMSE | MAE | $R^2$ | |
| Mordred | RF | 3.64 → ~3.56 | 3.46 → ~3.40 | −25.22 → ≈ | Error ↓ |
| | SVR | 3.76 → ~3.70 | 3.60 → ~3.55 | −26.98 → ≈ | Same pattern |
| | | | | | |
| MACCS | RF | 3.88 → 4.06 | 3.75 → 3.99 | −28.93 → −31.68 | No benefit |
| | SVR | 3.95 → ~3.96 | 3.79 → ~3.95 | −30.02 → ↓ | Failure persists |

Because Challenge Dataset 2 represents extreme experimental and chemical domain shift, **curation of the training data reduced noise** (lower RMSE/MAE) **but failed to recover calibration** (strongly negative $R^2$), demonstrating that curation cannot compensate for fundamental domain mismatch.

Curation improves reliability, not universality. Gains are observed only when descriptors are expressive enough (Mordred) and domain shift is moderate (Challenge 1). Under severe domain shift (Challenge 2), all models fail regardless of curation, indicating that representation and data mismatch dominate over noise. The challenge datasets therefore serve not as targets for improvement, but as diagnostic tools revealing model failure modes.

## 6 Conclusion

AqSolDB provides broad chemical coverage, but its protocol heterogeneity, hidden uncertainty, and domain mismatch issues limits its usefulness for drug discovery.

Our curation strategy, filtering to high-quality sources, weighting by reliability, and aligning with drug-like chemistry, shows measurable benefits: models trained on the curated set perform better on Challenge Dataset 1 when using expressive descriptors like Mordred. However, curation alone cannot overcome extreme domain shifts (as seen in Challenge Dataset 2) or compensate for weak representations (like MACCS).

This confirms that better data curation leads to better models, but only when the evaluation context is compatible. For real-world solubility prediction, investing in clean, well-documented, pharma-relevant data is more valuable than scaling up noisy benchmarks.

## 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. *Sci. Data* 6, (2019).

2. Chaudhry, S. & Khan, Z. *An Investigation into Aqueous Solubility*.

3. Fang, C. *et al.* Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *J. Chem. Inf. Model.* 63, 3263–3274 (2023).

4. ChEMBL documentation: *activities/assays; physicochemical assay type includes solubility-related measurements*.

5. Biogen Molecular Informatics repository (Computational-ADME): *public set of 3,521 commercially sourced compounds*; released ADME_public_set_3521.csv.

6. Polaris Hub dataset card (biogen/adme-fang-v1): *endpoint naming and benchmark context* (e.g., solubility at pH 6.8).