Metrics and Model Behavior - Beyond RMSE

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

So far, we evaluated our LogS models mainly with RMSE and R^2, treating them as final performance scores. For noisy, heterogeneous solubility data, these aggregates hide important questions: are predictions biased, are errors robust to outliers, and where do models fail when chemistry shifts? In this report, metrics are used as diagnostic tools to understand model behaviour, not just to rank models with the aim of understanding model behaviour, calibration, robustness, and failure regions. The central question is whether deeper evaluation leads to different conclusions than RMSE and R^2 alone, and how these insights can guide better model development.

Predicting aqueous solubility (LogS) relies on experimentally noisy and chemically diverse datasets, where protocol differences, solid-state effects, and varying conditions all introduce additional variability (Sorkun et al., 2019; Avdeef, 2019). In such a setting, the choice of evaluation metrics is not a minor technicality but central to obtaining a meaningful assessment of model performance. Metrics must allow us to distinguish between random experimental noise, systematic model bias, and genuine predictive failure.

Why this set of metrics?

To capture different facets of model behaviour, we first considered **Pearson correlation (r)**, **Concordance Correlation Coefficient (CCC)**, **distance correlation (dCor)**, **Information Coefficient (IC)**, **Mean Absolute Error (MAE)**, **Median Absolute Error (MedAE)**, and **Huber loss**. Correlation-based metrics describe how strongly predictions and experimental LogS values move together, whereas error-based metrics focus on the magnitude and robustness of the residuals. Each metric was selected to answer a specific diagnostic question about model behaviour rather than to optimise a single numerical score.

Pearson r measures linear association and is useful for assessing whether a model can rank compounds by solubility. However, it overlooks systematic bias and scale shifts, which frequently occur in solubility prediction. A model may therefore show high correlation while consistently over- or underestimating LogS. To address this limitation, the Concordance Correlation Coefficient (CCC) was therefore included to jointly evaluates correlation and agreement with the identity line, penalising constant offsets or rescaling and making it more suitable for assessing true predictive concordance (Lin, 1989).

Because solubility mechanisms can be non-linear, dCor was added to capture both linear and non-linear dependence. Comparing Pearson r and dCor allows us to determine whether poor performance arises from missing non-linear structure or from a fundamental lack of informative features. Information Coefficient (IC) was initially considered for similar reasons but was excluded from the final analysis because it largely duplicated the information captured by dCor while being less interpretable in this context.

To quantify prediction accuracy, error-based metrics were used. MAE is intuitive but sensitive to skewed residual distributions and outliers. Given that extreme deviations are expected in LogS data, MedianAE was

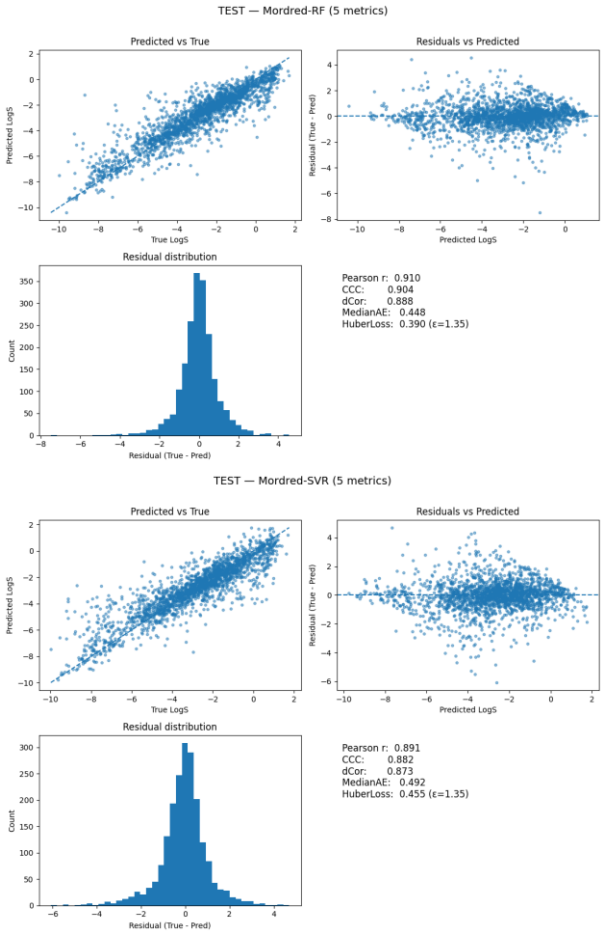
preferred as a robust measure of typical performance. Huber loss was included as a compromise between RMSE-like sensitivity to small errors and MAE-like robustness to large ones, allowing differentiation between global shifts and failures driven by isolated outliers (Oja et al., 2022).

Based on these considerations, the **final toolkit consists of CCC, Pearson r, dCor, MedianAE, and Huber loss**, providing information about agreement, association, robustness, and error magnitude that would remain hidden when relying on RMSE and R^2 alone.

In the following, these metrics are applied to Mordred- and MACCS-based RF/SVR models on the original test split and two challenge datasets. For each case, the metric profiles are used to diagnose three types of behaviour: in-distribution fit and calibration (test split), moderate domain shift with limited feature expressiveness (Challenge 1), and severe domain or target extrapolation accompanied by systematic bias (Challenge 2). These diagnoses are then contrasted with conclusions drawn from RMSE/R^2 alone, and implications for improving future solubility models are discussed.

With Mordred Descriptors

1.Test Split: Learned and Well-Calibrated
Pearson r (0.91 RF, 0.89 SVR), CCC (0.90, 0.88), and dCor (0.89, 0.87) are all high and closely aligned, indicating that both Mordred RF and SVR capture a strong, predominantly linear relationship between descriptors and LogS with minimal systematic bias. MedianAE values (0.45–0.49) and low Huber loss (0.39–0.46) show that typical errors are small and not dominated by outliers, with RF performing slightly more tightly than SVR. Together, this metric profile reflects a well calibrated, stable model operating within its training distribution.

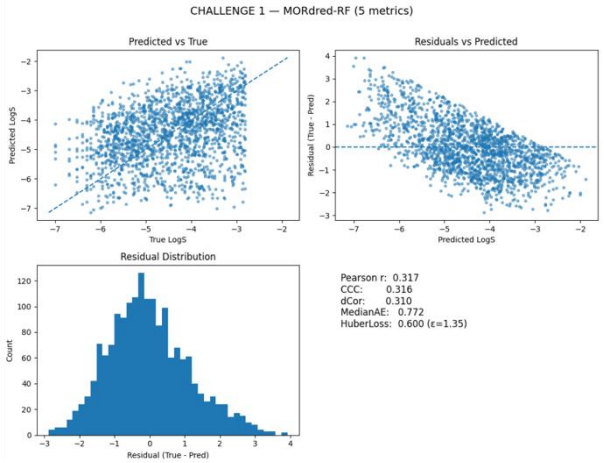


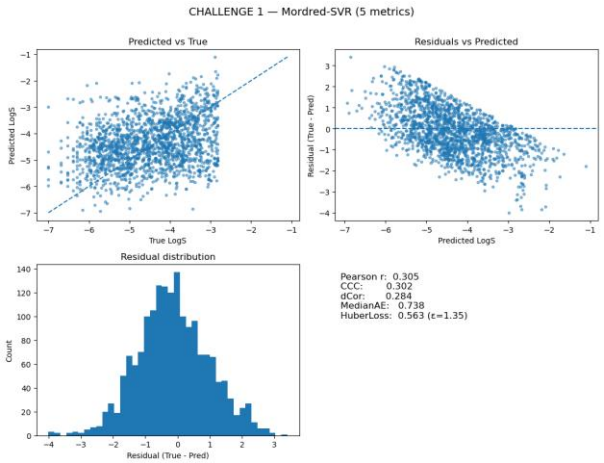
2. Challenge Data 1: Calibrated but Noisy

Pearson r (0.32) & dCor (0.31) are nearly identical. This confirms that the relationship between the Mordred descriptors and LogS remains largely linear but weaker than in the test split. Our model is capturing a genuine moderate signal. It can roughly sort molecules from soluble to insoluble, though it lacks precision ranking. **CCC (0.32)** closely matches Pearson r, suggesting no evidence of systematic bias within this dataset. **MedianAE (0.77)** indicates that for half of the dataset, our prediction is within a very narrow margin of the truth. **HuberLoss (0.60)** is exceptionally low, even lower than the MedianAE. A value this low indicates that our model's errors are **normally distributed** (a tight bell curve).

Taken together, this profile indicates an unbiased but noisy model: the general

solubility trend is captured, but predictions are scattered, reflecting limited precision under moderate domain shift.





3. **Challenge Data 2: Systematic Shift**

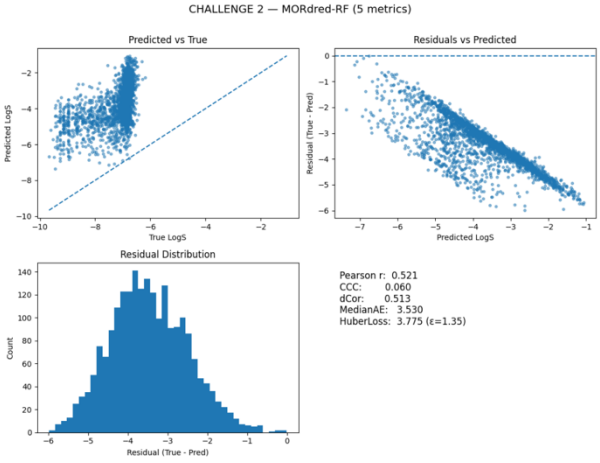
Pearson r (0.52) & dCor (0.51) are higher than in Challenge 1, indicating that Mordred descriptors still capture the relative ordering of solubility values. However, **CCC** drops sharply to **0.06**, revealing a severe systematic bias. The model learns the correct trend (slope) but predicts values that are consistently shifted from the true LogS scale.

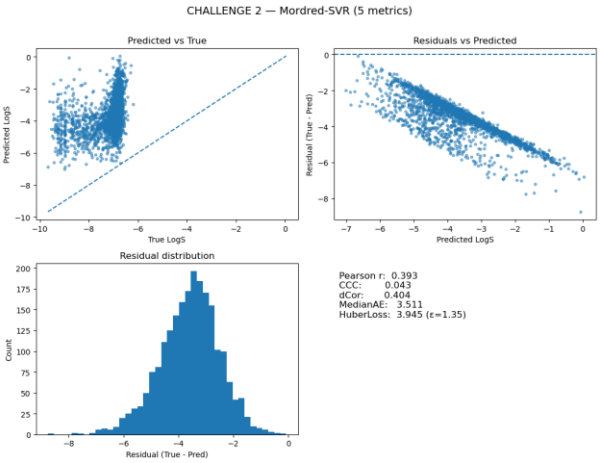
This failure is confirmed by the very large **MedianAE (3.53)** and **Huber loss (3.77)**. The similarity between these values indicates that the high error is not driven by a few outliers but reflects a global shift affecting nearly all predictions. Errors are large and consistent, rendering the model quantitatively unusable for this dataset.

4. Diagnosis of Failures

1. Challenge Data 1: Precision Limit

While the model is unbiased, it suffers from high variance. The Mordred descriptors capture the general chemical trend, but the model cannot pin down the exact value for specific molecules, resulting in a "cloud" of predictions rather than a tight line.





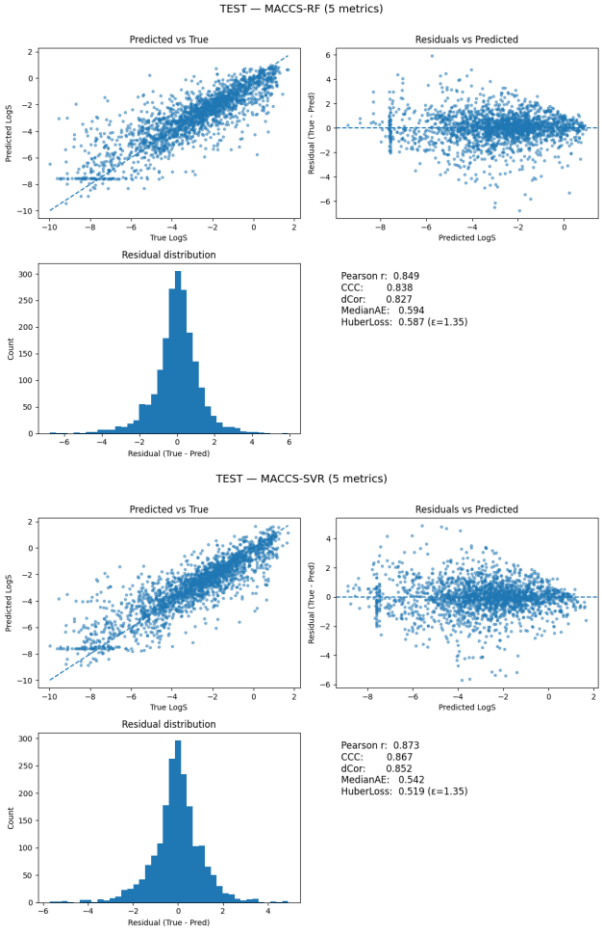
2. Challenge Data 2: Calibration Failure (Target Extrapolation)

The features successfully capture the ranking, but the model fails to predict the correct magnitudes. Because the training data likely lacked highly insoluble molecules (LogS < -6), the model hits a "floor" and physically cannot predict the extremely low values (e.g., -9 or -10) required for this set.

With MACCS Keys

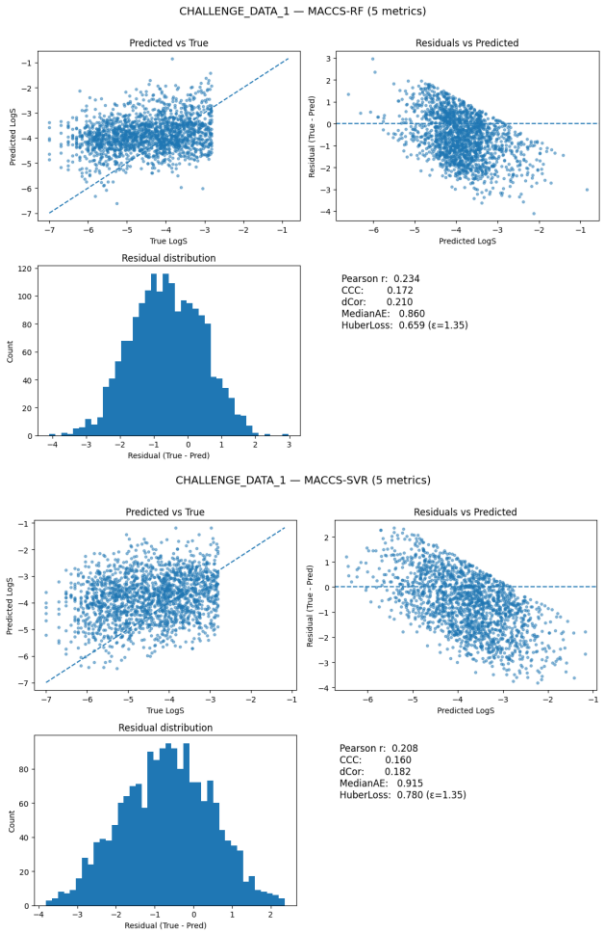
1. Test Split: Learned In-Distribution Pattern

Pearson r (0.85–0.87) and **dCor (0.83–0.85)** indicate strong dependence between MACCS features and LogS within the training chemical space. **CCC values (0.84–0.87)** closely match Pearson r, showing no evidence of systematic bias. **MedianAE (0.54–0.59)** is comparable to reported experimental uncertainty (≈0.5–0.7 log units), suggesting near-experimental accuracy. **Huber loss values (0.52–0.59)** further indicate stable performance, with SVR slightly smoother than RF. Overall, the model performs well when test molecules resemble the training distribution.



2. Challenge Data 1: Weak Ranking, Moderate Accuracy

Pearson r (0.21–0.23) and **dCor (0.20)** are low and nearly identical, indicating weak association and no hidden non-linear signal. **CCC (0.16–0.17)** confirms unreliable agreement. **MedianAE (0.86–0.91)** suggests moderate typical error, while **Huber loss (0.66–0.78)** indicates stable but consistently inaccurate predictions. This profile reflects limited feature expressiveness i.e, MACCS keys fail to capture the chemical distinctions required for ranking under domain shift.



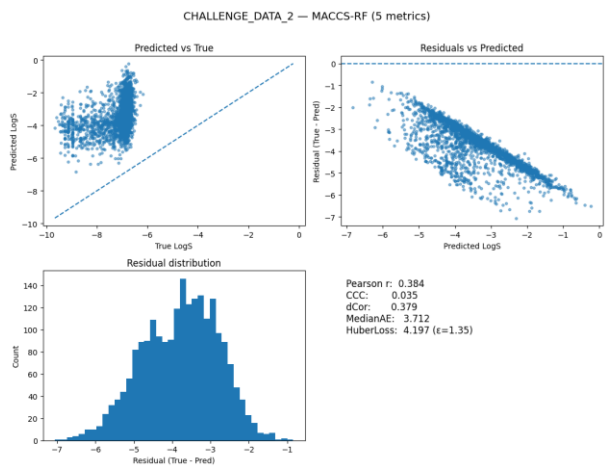
3. Challenge Data 2: Systematic Bias and Domain Failure

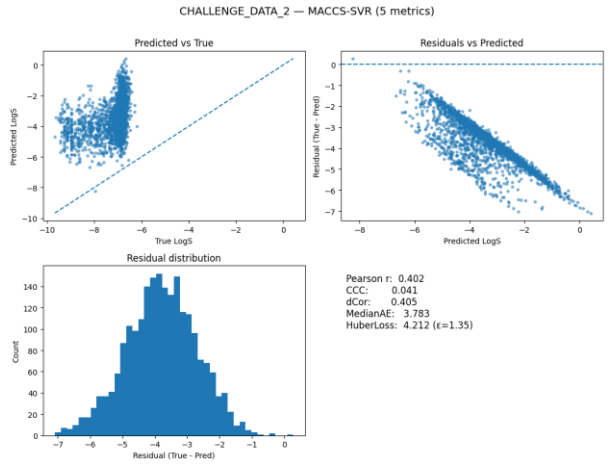
Pearson r (0.38 - 0.40) and dCor (0.40) indicate moderate ranking ability. The model is better at knowing which molecule is more soluble than another compared to Data 1, but **CCC (0.03 - 0.04)** is the critical warning **sign.** Pearson (0.4) combined with a near-zero CCC (0.04) is revealing severe systematic bias. Our model correctly identifies trends (X is more soluble than Y) but the values are shifted

significantly (e.g., the model predicts -2, but the reality is -6).

MedianAE (3.7 - 3.8) and **HuberLoss (4.19 - 4.21)** confirm large, consistent errors across the dataset, consistent with global miscalibration rather than isolated outliers. This confirms that the high error isn't just due to a few "weird" molecules. **The entire set of predictions is shifted.** The model is consistently missing the target by a wide margin (likely ~4 log units) for the majority of the dataset.

dCor (0.40): Again, dCor is almost identical to Pearson. This confirms the relationship is predominantly linear (or monotonic). The model clearly "sees" a signal here (better than in Data 1), as indicated by the moderate dependence (0.40).





4. Diagnosis of Failures

1. Challenge Data 1: Feature Expressiveness Limit.

MACCS keys are too simple. The model fails to differentiate between simple and complex versions of the same scaffold because the fingerprint does not account for the count or location of functional groups.

2. Challenge Data 2: Applicability Domain Failure.

These molecules are outside the chemical space our model understands. They are significantly more insoluble than our training data.

RMSE/R^2 vs Extended Metric Perspective

Using RMSE and R^2 alone, both challenge datasets would simply appear “harder” than the test split. However, the extended metric analysis reveals qualitatively different failure modes. **Challenge Data 1 reflects limited feature expressiveness and increased noise with preserved calibration**, while **Challenge Data 2 represents severe extrapolation and systematic bias**. These distinctions are invisible to RMSE/R^2 but are critical for deciding how to improve the models.

Summary

Taken together, the extended metrics show that Mordred based models are strong and well calibrated in distribution but primarily fail due to precision limits and target extrapolation, whereas MACCS based models suffer from both feature expressiveness limits and applicability domain failure under domain shift. These patterns would be blurred if only RMSE and R^2 were reported. The deeper

analysis reveals why models fail, not just that they fail, and points directly to improvement strategies such as expanding the LogS range of training data, enforcing applicability domain checks, recalibrating predictions under shift, and adopting richer, physically informed molecular representations.

References

1. Avdeef, A. (2019). Solubility of ionizable drugs. ADMET & DMPK, 7(1), 13–38.
2. Lin, L. I. K. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1), 255–268.
3. Oja, M., Kontsedalov, A., Tamm, T., & Maran, U. (2022). Intrinsic aqueous solubility: Mechanistically transparent data-driven modeling of drug substances. Pharmaceutics, 14(11), 2248.
4. Sorkun, M. C., Koşka, A., & Er, S. (2019). AqSolDB: A curated reference set of aqueous solubility data. Scientific Data, 6, 143.