Task 04: Metrics and Model Behaviour - Beyond RMSE

Title: Not All Errors Are Equal

Baseline Model and Motivation

So far we trained a baseline ligand-based QSAR regression model to predict binding affinity–related values for dopamine D_2 receptor (DRD2) ligands. The dataset consisted of small molecules collected from different bioactivity databases where each molecule was represented using two-dimensional molecular features. These features included Morgan (ECFP) fingerprints which encode local atomic environments, and classical RDKit descriptors that describe physicochemical properties such as molecular weight, lipophilicity, polarity and hydrogen-bonding capacity. The target variable was treated as a continuous affinity value derived from reported Ki or IC50-related measurements.

The model performance was initially evaluated using standard regression metrics namely the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). RMSE was used to measure the average magnitude of prediction errors while R^2 was used to quantify how much of the variance in the target values could be explained by the model. Based on these metrics the model showed reasonable performance on the test set and these values were reported as the baseline results.

However it became clear that RMSE and R^2 alone provide only a limited understanding of model behavior. RMSE summarizes all prediction errors into a single value and gives higher weight to larger errors but it does not show where these errors occur or whether they are systematic. Similarly R^2 reflects global variance explained but does not indicate whether the model performs equally well across different regions of chemical space.

This limitation is especially important in the context of binding affinity prediction where the labels themselves are known to be noisy and assay-dependent. As discussed in Task 2, affinity values collected from heterogeneous experimental setups contain irreducible uncertainty which may strongly affect regression performance. In such cases two models with similar RMSE and R^2 values may still behave very differently when examined in more detail.

For these reasons, RMSE and R^2 should be considered baseline summary metrics rather than sufficient indicators of model quality. They motivate the need for additional evaluation methods that analyze errors more explicitly and help reveal how and where the model fails.

Evaluation Metrics Beyond RMSE in log Ki Prediction

In the evaluation of regression models for binding affinity log Ki prediction, the Root Mean Square Error (RMSE) is often used but can be misleading due to its sensitivity to outliers [1]. In computational chemistry, datasets frequently contain activity cliffs, where minor structural changes in a SMILES string result in significant jumps in potency, leading to non-Gaussian error distributions [2][3]. Because RMSE squares the residuals before averaging, a handful of poor predictions on rare chemical scaffolds can

disproportionately inflate the error score, masking the performance of a model that is otherwise accurate for the majority of the chemical space [1][4].

To achieve a more robust assessment, the Mean Absolute Error (MAE) and Median Absolute Error (MedAE) offer critical alternatives. MAE provides an intuitive, linear measure of the typical error in log units, treating all deviations with equal weight [1]. For datasets with extreme outliers, MedAE is even more resilient. According to the PerMetrics Documentation, MedAE is particularly valuable because it focuses on the central tendency of the error distribution by calculating the median of absolute differences predicted y value and actual y value, effectively filtering out the influence of catastrophic misses on anomalous chemical motifs [5]. By comparing the gap between RMSE and MedAE, researchers can diagnose whether a model's error is uniform or driven by specific failure regions. [3][4].

For models utilizing Morgan Fingerprints and architectures like XGBoost or MLP, these metrics are essential for defining the Applicability Domain. Beyond aggregate numbers, analyzing residuals against descriptors such as molecular weight or lipophilicity allows for the identification of systematic biases that a single value cannot capture [6]. This diagnostic approach ensures that log Ki predictions are understood within the structural and chemical limitations of the model, rather than being taken for granted based on a single global score [3][6].

Residual & Error Distribution Analysis

To rigorously evaluate the model's reliability, we conducted a residual analysis based on the fundamental definition: Residual = $y - \hat{y}$ where y is the experimental value and $\hat{y}$ is the predicted value. This analysis aims to differentiate between irreducible random noise and systematic modeling errors and focuses on three key diagnostic criteria: randomness, bias and homoscedasticity.

1. Residual Patterns and Diagnostics: The analysis revealed a sharp contrast in model behavior between the two datasets:

Internal Test Set (Homoscedasticity): The residuals displayed a symmetric and random distribution centered around zero. This lack of discernible structure indicates that the model has successfully captured the deterministic physicochemical information available in the training data and leaves only Gaussian white noise.

Challenge Set (Systematic Bias): On the other hand, the Challenge Set exhibited a distinct and non-random pattern. The residuals were not independent; rather they demonstrated a systematic bias dependent on the actual values. Specifically the model tended to predict values close to the training mean regardless of the actual affinity. According to regression diagnostics principles, such a structured pattern is a definitive indicator of model deficiency and proves that the model failed to generalize the underlying rules to the novel scaffolds [7].

2. Error Distribution Characteristics: We further quantified the generalization gap by analyzing the shape of the error distributions:

Internal Consistency: The error distribution followed a narrow and normal (Gaussian) curve. This confirms that the model is unbiased and precise within its applicability domain. On the other hand the error distribution for the Challenge Set deviated significantly from normality and appeared flatter with a much wider variance. This confirms that the high RMSE observed is not due to random fluctuations but due to the model's inability to distinguish extremely active or inactive compounds in new chemical spaces. The disparity between the narrow, Gaussian errors of the internal set and the wide, structured errors of the challenge set provides strong statistical evidence of overfitting. The model exhibits low bias on training data but high variance on external data. This behavior confirms that the model is memorizing the training set patterns rather than generalized features, rendering it unreliable for the external Challenge Set without further optimization [8].

Failure Regions and Model Generalization Beyond Aggregate Metrics

Evaluation metrics such as RMSE and R^2 are widely used to summarize the performance of regression models in cheminformatics; however, these metrics can provide a misleading sense of confidence when models are evaluated using standard train-test splits [9]. It has been shown that many such splits contain structurally similar molecules in both training and test sets, allowing models to perform well by learning local patterns rather than developing transferable structure-activity relationships [10]. Consequently, strong test set performance does not necessarily indicate that a model will remain reliable when applied to chemically novel compounds, particularly those lying outside the training distribution [11].

This limitation is commonly explained through the concept of **failure regions**, which describe areas of chemical space where the model has limited training exposure and therefore struggles to make accurate predictions [12]. Although Morgan fingerprints and descriptors are effective at encoding features present in the training data, they do not guarantee robust behavior when the model encounters unfamiliar scaffolds or extreme physicochemical properties [13]. In such cases, prediction errors are not randomly distributed but tend to cluster around molecules that are chemically distant from the training set, leading to systematic rather than isolated failures [9][14].

The concept of failure regions is especially relevant in drug discovery, where predictive models are often used to evaluate compounds that differ substantially from those seen during training, such as in scaffold hopping or lead optimization efforts [15]. In these situations, even small structural changes can cause large shifts in binding affinity, a phenomenon known as activity cliffs, which ligand based machine learning models commonly find difficult to capture [11][14]. As a result, models that appear reliable within familiar chemical space may still produce misleading predictions when applied to new and chemically distinct molecules. This highlights the importance of moving beyond single-number performance metrics and instead examining how prediction errors change as molecular novelty increases [10]. By identifying failure regions, it becomes possible to better define a model’s applicability domain and to make more informed judgments about when its predictions can be trusted [12][15].

In our task, the model performs well on the test dataset but shows larger errors on the challenge dataset, which contains more chemically novel compounds. These molecules often have unfamiliar scaffolds or extreme physicochemical properties compared to the training data. This difference in performance highlights failure regions where the model must extrapolate beyond its learned chemical space.

Metrics Vs Behaviour Comparison

The evaluation results show that relying solely on global regression metrics such as RMSE and R^2 provides an incomplete assessment of QSAR model performance. Although these metrics indicate acceptable average error and variance explanation, they do not describe how prediction errors are distributed across chemical space [16]. RMSE is particularly sensitive to large deviations, while R^2 summarizes overall trends without revealing localized or systematic failures. In ligand-based binding affinity prediction, where experimental measurements are often noisy and assay-dependent, such aggregate metrics can obscure important weaknesses [17].

As a result, models with similar RMSE and R^2 values may exhibit substantially different predictive reliability when applied beyond familiar chemical regions [18]. A more informative interpretation emerges when robust metrics and residual-based diagnostics are considered. Metrics such as MAE and MedAE complement RMSE by providing linear and median-based estimates of typical prediction error, thereby reducing the influence of extreme outliers [19].

Differences observed between RMSE and MedAE indicate that prediction errors are not uniformly distributed but are driven by specific subsets of compounds. This conclusion is further supported by residual and error distribution analysis, where structured and non random patterns point to overfitting and limited generalization [20].

Further evidence of limited generalization is provided by failure region analysis. Despite strong performance on chemically familiar compounds, the model shows substantially larger errors on chemically novel molecules characterized by unfamiliar scaffolds or extreme physicochemical properties [17]. However, previous studies have shown that ligand-based machine learning models are particularly sensitive to activity cliffs and scaffold novelty, where small structural changes can lead to large shifts in binding affinity. Identifying such failure regions is therefore essential for defining the applicability domain and for determining when model predictions can be trusted.

Model Improvement:

To address specific failure modes we propose three modifications to the machine learning pipeline:

Switching to Ranking Loss: Since the training labels contain irreducible assay noise [21] and the model shows strength in relative ordering, we should replace Mean Squared Error (MSE) with a Pairwise Ranking Loss (e.g., RankNet). This forces the model to learn "Is Drug A > Drug B?" which is robust against experimental noise rather than trying to fit exact noisy numbers.

Scaffold-Aware Validation: The failure on the Challenge Set (Domain Shift) indicates our random split over-estimated performance. We propose adopting Bemis–Murcko scaffold splitting [22] to force the model to learn transferable structural rules rather than memorizing local chemical neighborhoods.

Uncertainty Quantification: Failure analysis showed the model struggles with scaffolds sparse in the training set. We suggest implementing Ensemble-based Uncertainty Estimation (e.g., Random Forest variance or Monte Carlo Dropout) to output a Confidence Score. This allows us to filter out unreliable

predictions that fall outside the model's applicability domain and directly addresses the systematic errors observed on novel scaffolds [23].

References

[1] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79-82.

[2] Sheridan, R. P. (2013). The effect of outliers on the performance of QSAR models. *Journal of Chemical Information and Modeling*, 53(11), 2837-2850.

[3] Pernot, P., et al. (2020). Impact of non-normal error distributions on the benchmarking and ranking of Quantum Machine Learning models. *Machine Learning: Science and Technology*, 1(3), 035011.

[4] Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14), 5481-5487.

[5] PerMetrics Documentation. "MedAE – Median Absolute Error." *Regression Metrics Guide*. [Online]. Available: permetrics.readthedocs.io

[6] Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6-7), 476-488.

[7] Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: John Wiley & Sons. (Chapter 2: Examination of Residuals).

[8] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. (Section 7.2: Bias, Variance and Model Complexity).

[9] Sheridan, R. P. (2013). The effect of outliers on the performance of QSAR models. *Journal of Chemical Information and Modeling*, 53(11), 2837–2850.

[10] Wu, Z., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9, 513–530.

[11] Pernot, P., et al. (2020). Impact of non-normal error distributions on benchmarking ML models. *Machine Learning: Science and Technology*, 1(3), 035011.

[12] Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, 29(6–7), 476–488.

[13] Yang, K., et al. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8), 3370–3388.

[14] Kramer, C., et al. (2012). Multivariate models outperform univariate models in activity cliff prediction. *Journal of Chemical Information and Modeling*, 52(11), 2813–2822.

[15] Sheridan, R. P., & Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today*, 7(17), 903–911.

[16] B. Chen *et al.*, “Universities of Leeds , Sheffield and York Evaluation of Machine-Learning Methods for Ligand-Based Virtual Screening,” vol. 21, pp. 53–62, 2007.

[17] K. Yang *et al.*, “Analyzing Learned Molecular Representations for Property Prediction,” 2019, doi: 10.1021/acs.jcim.9b00237.

[18] C. Res, C. J. Willmott, and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” vol. 30, pp. 79–82, 2005.

[19] P. Gramatica, “Principles of QSAR models validation : internal and external,” no. 5, pp. 694–701, 2007, doi: 10.1002/qsar.200610151.

[20] T. Chai, R. R. Draxler, and C. Prediction, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” no. 2005, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[21] G. A. Landrum and S. Riniker, "Combining IC50 or Ki values from different sources is a source of significant noise," *J. Chem. Inf. Model.*, vol. 64, no. 5, pp. 1560–1567, 2024.

[22] Z. Wu et al., "MoleculeNet: a benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, pp. 513–530, 2018.

[23] B. Lakshminarayanan et al., "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," *NeurIPS*, 2017.