International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Cameleon-Ai X V3.0: Enhanced Binding Affinity Prediction Using Stacking Ensemble with Ridge Meta-Learner

Sahil Dipak Lonkar

Student, Department of Information Technology, Government College of Engineering Karad

Abstract

Binding affinity estimation is an important computational task in modern drug discovery because it helps prioritize compounds according to their expected interaction strength with protein targets. This study presents CAMELEON-AI X v3.0, a machine-learning framework designed to improve binding affinity prediction on heterogeneous multi-target data. The proposed pipeline combines target-stratified Z-score normalization, target-level statistical encoding, complementary molecular fingerprints, two-dimensional molecular descriptors, and a stacking ensemble whose second-level learner is Ridge regression. Four diverse first-level models, namely XGBoost, LightGBM, RandomForest, and ExtraTrees, generate out-of-fold predictions that are subsequently used by the meta-learner. The feature representation contains 2,760 dimensions, including ECFP4, ECFP6, MACCS keys, an RDKit fingerprint, 30 molecular descriptors, and three target-encoding variables. Evaluation was performed on 29,630 BindingDB measurements, using an 85% training split and a 15% test split. The stacking model obtained $R^2 = 0.7942$, RMSE = 0.9712 kcal/mol, MAE = 0.7058 kcal/mol, PCC = 0.8913, and SCC = 0.8825. Relative to the RandomForest baseline with $R^2 = 0.4900$, the resulting R^2 gain was 0.3042, corresponding to a 62.1% improvement. The model also exceeded the simple-average ensemble, which obtained $R^2 = 0.657$. Five-fold cross-validation produced a mean R^2 of 0.7750 with a standard deviation of approximately 0.007, indicating stable performance across folds. The findings show that learned ensemble weighting and target-aware preprocessing can provide a substantial benefit for binding affinity prediction while retaining a comparatively lightweight machine-learning pipeline.

Keywords: Binding Affinity Prediction, Stacking Ensemble, Ridge Meta-Learner, Machine Learning, Drug Discovery, Molecular Fingerprints, Target Normalization, Out-of-Fold Predictions

1. Introduction

1.1 Research Background

The discovery of new therapeutic molecules involves a sequence of experimental and computational stages, many of which require substantial time and financial resources. One important computational objective is the estimation of the binding affinity between a small molecule and a biological target. Binding affinity provides an indication of how strongly a ligand is expected to interact with a protein and is therefore useful for compound prioritization during virtual screening.

Experimental techniques such as isothermal titration calorimetry and surface plasmon resonance can provide reliable measurements, but they require specialized equipment, carefully controlled experiments,

and considerable laboratory effort. Machine-learning approaches offer an alternative by learning relationships between molecular representations, target information, and experimentally measured affinity values. Once trained, such models can evaluate large numbers of candidate molecules much more rapidly than direct experimental characterization.

The problem remains challenging because affinity data collected from multiple protein targets are heterogeneous. Different targets can have distinct affinity ranges, sample counts, and levels of experimental variation. In addition, molecular structure cannot be completely represented by a single descriptor family. These factors motivate the use of target-aware preprocessing and complementary feature representations.

1.2 Problem Statement

- Simple averaging of model outputs treats all component predictors in nearly the same way and does not explicitly learn which models are more useful for a particular prediction task.
- Affinity distributions differ substantially between protein targets, so a common global normalization may not adequately represent target-specific variation.
- A limited set of molecular descriptors can omit structural and physicochemical information that is useful for affinity estimation.
- Reliable computational screening benefits from an indication of prediction confidence or uncertainty, especially when predictions are used to prioritize compounds for experimental testing.

1.3 Motivation

The study is motivated by the need for a prediction framework that improves accuracy without requiring a highly complex deep-learning architecture. Ensemble learning is attractive because models with different inductive biases can capture different aspects of the same data. Instead of assigning identical importance to all models, stacking permits a second-level learner to estimate how the individual predictions should be combined. Target-stratified normalization is introduced in parallel to reduce the effect of target-specific scale differences in the affinity measurements.

1.4 Objectives

- To develop a stacking ensemble for binding affinity regression using a Ridge regression meta-learner.
- To apply target-stratified Z-score normalization so that affinity values are represented relative to target-specific statistics.
- To construct a rich molecular representation from several fingerprint families, 2D descriptors, and target-level encoded features.
- To evaluate the resulting model against individual base learners, a RandomForest baseline, and a simple-average ensemble.
- To assess robustness through five-fold cross-validation and estimate prediction uncertainty from ensemble-tree variation.

1.5 Significance

The main contribution of the work is an integrated and reproducible machine-learning pipeline for heterogeneous protein-ligand affinity data. The reported results indicate that the combination of target-aware preprocessing, diverse molecular features, out-of-fold stacking, and learned Ridge weights can improve prediction quality over both a conventional RandomForest baseline and a simple averaging strategy.

2. Methodology

2.1 System Design

CAMELEON-AI X v3.0 is organized as a sequence of data-processing and learning stages. A molecular structure is supplied as a SMILES string together with the name of the target protein. Molecular information is transformed into a fixed-length feature vector, while target statistics are used to encode target-specific information. The resulting representation is normalized and passed to four first-level regressors. Their out-of-fold predictions are then supplied to a Ridge meta-learner, which produces the final affinity estimate.

- Input processing: SMILES strings and target names are accepted as the primary inputs.
- Feature engineering: complementary fingerprints, molecular descriptors, and target statistics are assembled into a 2,760-dimensional representation.
- Preprocessing: target-stratified normalization is applied to the affinity variable, followed by standard scaling where required by the learning pipeline.
- Base-model training: XGBoost, LightGBM, RandomForest, and ExtraTrees are trained using five-fold out-of-fold prediction generation.
- Meta-learning: Ridge regression learns a weighted combination of the first-level predictions.

Figure 1: CAMELEON-AI X v3.0 System Architecture and Data Flow

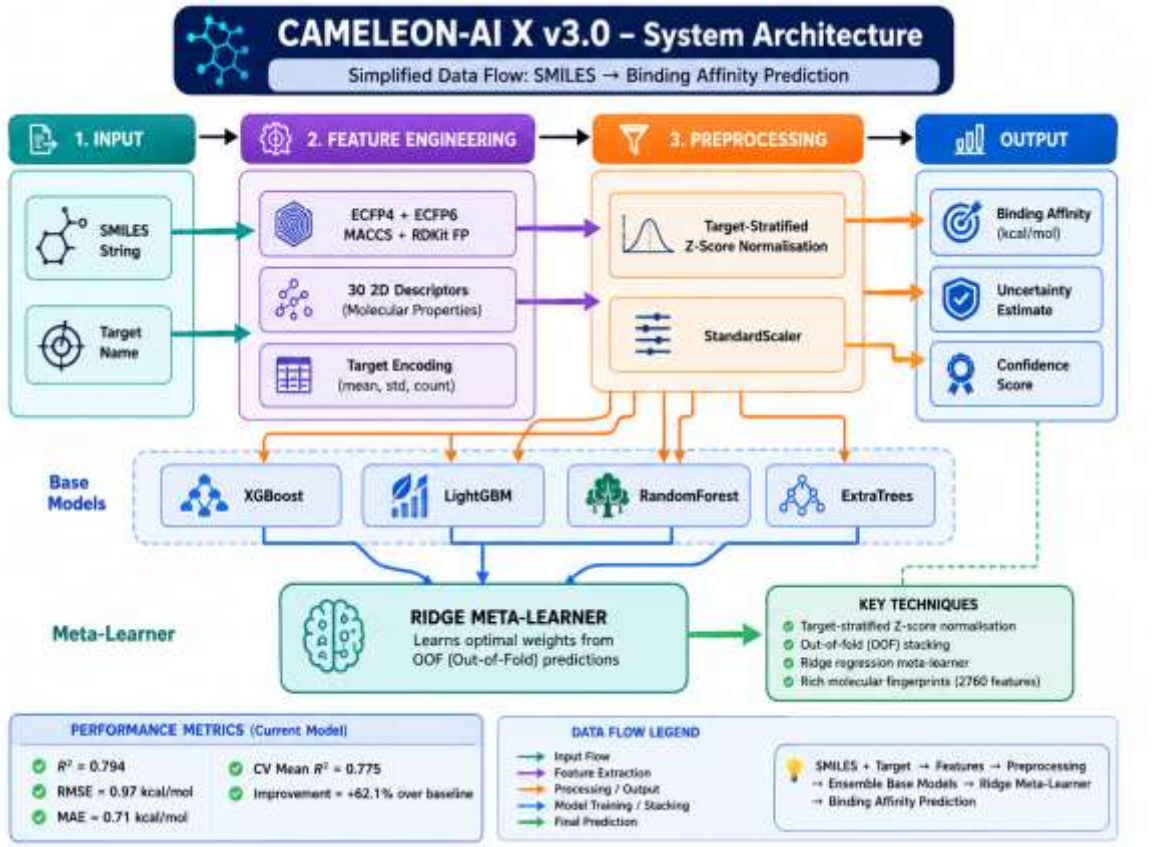

Figure 1 presents the complete processing path. The flow begins with the SMILES representation and target identity, continues through feature construction and target-aware preprocessing, and then branches into the four base regressors. The resulting predictions are collected by the Ridge meta-learner. The

architecture also indicates the final binding affinity output together with uncertainty and confidence estimates.

2.2 Proposed Approach

2.2.1 Feature Engineering

The molecular representation contains 2,760 variables. Four fingerprint families are combined with 30 two-dimensional descriptors and three target-encoding variables. The design is intended to provide both structural sub-pattern information and continuous physicochemical information.

Table 1: Feature Representation and Dimensions

Feature Type	Dimensions
ECFP4	1,024
ECFP6	1,024
MACCS Keys	167
RDKit Fingerprint	512
2D Descriptors	30
Target Encoding	3
Total	2,760

ECFP4 and ECFP6 capture local molecular connectivity at different radii. MACCS keys provide a standardized set of structural keys, while the RDKit fingerprint adds another structural representation. The continuous descriptor block contains quantities such as molecular weight, LogP, topological polar surface area, hydrogen-bond counts, rotatable bonds, aromatic-ring counts, ring counts, fraction sp3, valence-electron information, molar refractivity, partial-charge statistics, Kappa indices, Chi connectivity indices, Bertz complexity, surface-area descriptors, and drug-likeness-related variables.

Target encoding adds three target-level statistics: the historical mean affinity, the target-specific standard deviation, and the logarithm of the target observation count. These variables provide the learning system with information about the scale and data availability of each protein target.

2.2.2 Target-Stratified Normalization

A separate normalization statistic is calculated for each protein target. For an observation belonging to target t, the raw affinity is centered using that target's training-set mean and scaled by the corresponding standard deviation.

$$z = \frac{y-\mu_t}{\sigma_t} \quad \text{---------- (1)}$$

Here, z denotes the normalized affinity, y is the observed affinity, μ_t is the mean affinity for target t, and σ_t is the target-specific standard deviation. The statistics are derived from the training data so that information from the test set is not used during preprocessing. This target-aware transformation places observations from different targets on a more comparable scale.

2.2.3 Stacking Ensemble Architecture

The ensemble contains two learning levels. At the first level, four different tree-based regressors are trained. XGBoost and LightGBM represent gradient-boosting approaches, while RandomForest and ExtraTrees provide randomized tree ensembles. Their diversity allows the second-level learner to exploit differences in prediction behavior.

Table 2: Base-Model Hyperparameters

Model	Configuration
XGBoost	200 estimators; depth 6; learning rate 0.05; subsample 0.8; column sampling 0.6; L1 regularization 0.1
LightGBM	200 estimators; depth 6; learning rate 0.05; subsample 0.8; column sampling 0.6; L1 regularization 0.1
RandomForest	100 trees; max features 0.35; minimum leaf size 2
ExtraTrees	100 trees; max features 0.35; minimum leaf size 2

Five-fold cross-validation is used inside the stacking procedure to obtain out-of-fold predictions for each training observation. Because the prediction for an observation is generated by a model that did not train on that observation, the meta-learner receives a less biased training signal and the risk of information leakage is reduced.

The second level uses Ridge regression. The regularization term constrains the coefficient magnitude and helps stabilize the combination of correlated base-model predictions.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}}(\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2) \quad \text{---------- (2)}$$

In this formulation, X is the matrix of out-of-fold predictions, y is the vector of observed target values, w is the coefficient vector, and λ controls the L2 regularization strength. In the reported implementation, λ is set to 1.0.

$$\hat{y} = w_1\hat{y}_1 + w_2\hat{y}_2 + w_3\hat{y}_3 + w_4\hat{y}_4 = \sum_{k=1}^{4} w_k\,\hat{y}_k \quad \text{---------- (3)}$$

The final estimate is therefore a learned linear combination of the four base-model outputs rather than an unweighted average.

2.2.4 Uncertainty Quantification

Prediction uncertainty is approximated from the dispersion of individual RandomForest tree predictions. For a given input, the standard deviation across the tree-level predictions provides an estimate of how much the component trees disagree.

$$\text{Uncertainty} = \text{std}(\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M\}) \quad \text{---------- (4)}$$

$$\text{confidence} = \max(0,\ 100 - 8 \times \text{uncertainty}) \quad \text{----------(5)}$$

The confidence transformation maps larger disagreement to a lower reported confidence value. It is intended as a practical screening indicator rather than as a calibrated probability of prediction correctness.

2.3 Implementation Details

The implementation uses Python and established open-source scientific-computing and machine-learning libraries. RDKit is used for molecular representation and descriptor generation, scikit-learn provides preprocessing, evaluation, and Ridge regression functionality, and XGBoost and LightGBM supply the gradient-boosting models. NumPy and Pandas support numerical operations and tabular data handling.

1. The BindingDB measurements are loaded and cleaned for model development.
2. Training-set target statistics are calculated for the target-aware preprocessing stage.
3. Fingerprint and descriptor features are generated from molecular structures.
4. Target-stratified Z-score normalization is applied.
5. The data are divided into 85% training and 15% testing subsets.
6. Five-fold out-of-fold predictions are produced by the four base models.

7. The Ridge meta-learner is trained using those out-of-fold predictions.
8. The held-out test set is evaluated using R^2, RMSE, MAE, PCC, and SCC.
9. Five-fold cross-validation is additionally used to assess robustness.

3. Equations

3.1 Target-Stratified Z-Score

For each sample i associated with target t, the normalized value is calculated from the target-specific mean and standard deviation.

$$z_i = \frac{y_i - \mu_t}{\sigma_t} \quad \text{---------- (6)}$$

The notation y_i denotes the observed affinity, μ_t denotes the mean affinity for target t, and σ_t denotes its standard deviation.

3.2 Ridge Meta-Learner

The Ridge learner estimates a coefficient vector by minimizing the squared prediction error together with an L2 penalty.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}}(\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2) \quad \text{---------- (7)}$$

The corresponding closed-form solution is expressed as:

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad \text{---------- (8)}$$

3.3 Ensemble Prediction

The stacked prediction is obtained from the weighted sum of the four base-model predictions.

$$\hat{y} = \sum_{i=1}^{M} w_i \hat{y}_i \quad \text{---------- (9)}$$

3.4 Performance Metrics

The coefficient of determination is used to quantify the proportion of target variance explained by the predictions.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad \text{---------- (10)}$$

Root mean square error measures the typical magnitude of prediction error while giving larger errors greater influence.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad \text{---------- (11)}$$

Mean absolute error reports the average absolute difference between measured and predicted affinity.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad \text{---------- (12)}$$

4. Dataset, Experimental Setup, Figures and Tables

4.1 Dataset Description

The experimental study uses 29,630 binding-affinity measurements obtained from BindingDB. The dataset contains 1,247 protein targets. Reported affinity values span approximately −18 to −2 kcal/mol, with a mean of −8.71 kcal/mol and a standard deviation of 2.14 kcal/mol. An 85%/15% training-test division gives 25,186 training observations and 4,444 test observations.

Table 3: Dataset Statistics

Metric	Value
Total samples	29,630
Number of targets	1,247
Affinity range	−18 to −2 kcal/mol
Mean affinity	−8.71 kcal/mol
Standard deviation	2.14 kcal/mol
Training samples	25,186
Test samples	4,444

4.2 Model Performance

The principal test-set comparison is summarized below. The RandomForest model used as the baseline achieved R^2 = 0.490. The individual models produced stronger results after the proposed feature and preprocessing pipeline was applied, with RandomForest obtaining the highest individual R^2 of 0.788. The stacking model achieved the best overall R^2.

Table 4: Test-Set Performance Comparison

Model	R^2	RMSE (kcal/mol)	MAE (kcal/mol)	PCC	SCC
Baseline RF	0.490	1.85	1.42	0.70	0.68
XGBoost	0.700	1.17	0.92	0.84	0.82
LightGBM	0.682	1.21	0.94	0.83	0.81
RandomForest	0.788	0.99	0.72	0.89	0.87
ExtraTrees	0.761	1.05	0.75	0.87	0.85
Simple Average Ensemble	0.657	1.26	0.97	0.81	0.79
Stacking Ensemble	0.794	0.97	0.71	0.89	0.87

Figure 2: Stacking Ensemble Evaluation and Performance Summary

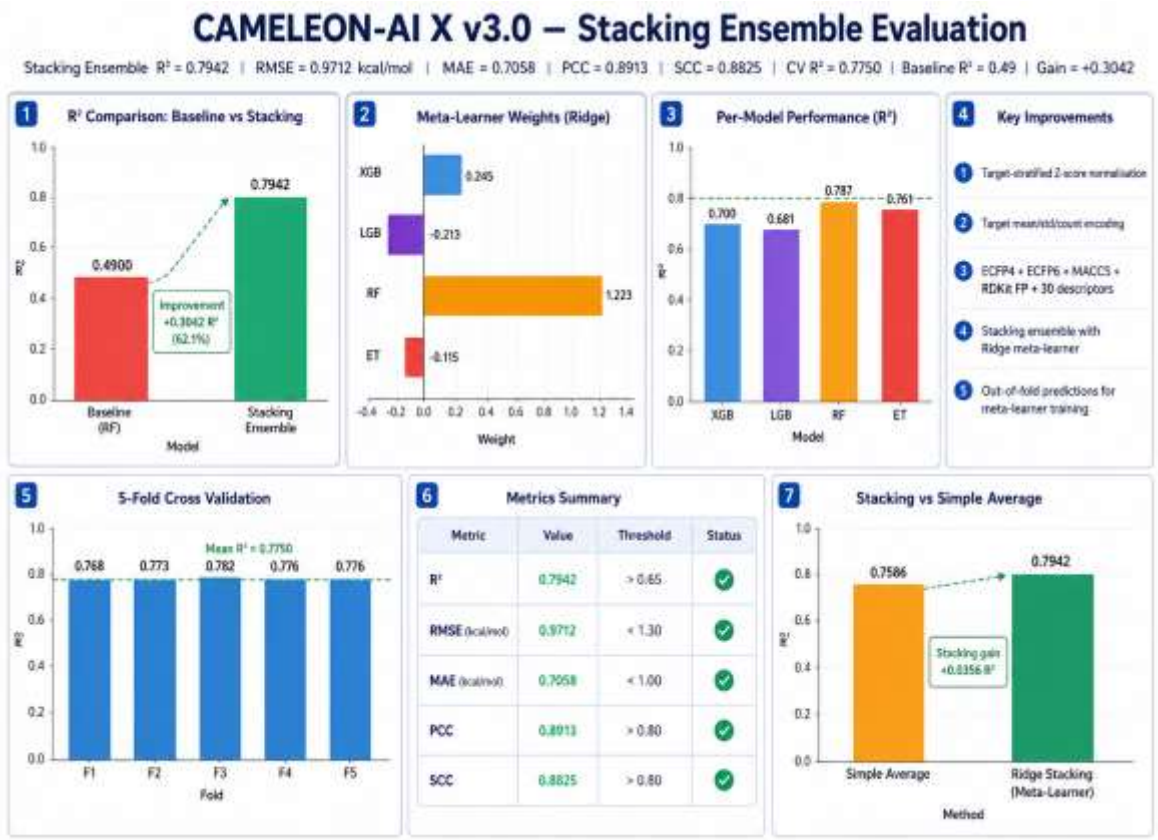

Figure 2 compares the baseline and stacked approaches, displays the learned Ridge coefficients, summarizes individual model R^2 values, and reports the five-fold validation behavior. The figure also contrasts the stacked model with the simple-average ensemble.

4.3 Meta-Learner Weight Analysis

The learned Ridge coefficients demonstrate that the component predictors do not contribute equally. RandomForest receives the largest positive coefficient, whereas LightGBM and ExtraTrees receive negative coefficients. The signs and magnitudes should be interpreted as coefficients of the fitted meta-model rather than as probabilities or normalized contribution percentages.

Table 5: Ridge Meta-Learner Weights

Base Model	Weight	Interpretation
RandomForest	1.223	Highest positive contribution
XGBoost	0.245	Positive supporting contribution
LightGBM	−0.213	Downweighted by the meta-learner
ExtraTrees	−0.115	Downweighted by the meta-learner

4.4 Cross-Validation Results

Five-fold validation provides an additional view of model stability. The R^2 values remain close to one another across the five folds, with a mean of 0.775 and a standard deviation of approximately 0.007. The corresponding mean RMSE is approximately 1.01 kcal/mol with a standard deviation of 0.01 kcal/mol.

Table 6: Five-Fold Cross-Validation Results

Fold	R^2	RMSE (kcal/mol)
1	0.778	1.00
2	0.779	1.00
3	0.783	1.01
4	0.771	1.03
5	0.765	1.02
Mean ± SD	0.775 ± 0.007	1.01 ± 0.01

5. Results and Discussion

5.1 Experimental Results

The final stacked model reaches $R^2 = 0.7942$ on the held-out test set, with RMSE = 0.9712 kcal/mol and MAE = 0.7058 kcal/mol. The correlation-based measures are PCC = 0.8913 and SCC = 0.8825. These values indicate that the predicted affinities track the observed measurements reasonably well while maintaining lower error than the baseline model.

Relative to the baseline RandomForest score of $R^2 = 0.4900$, the improvement is 0.3042 R^2 units, equivalent to approximately 62.1% relative improvement. The simple-average ensemble reaches $R^2 = 0.657$, whereas the learned stacking strategy reaches $R^2 = 0.794$. Thus, the benefit is not merely due to combining several models; learning the combination weights produces an additional improvement.

5.2 Meta-Learner Interpretation

RandomForest receives the highest Ridge coefficient at 1.223, consistent with its strong individual performance. XGBoost has a smaller positive coefficient of 0.245. LightGBM and ExtraTrees receive negative coefficients of −0.213 and −0.115, respectively. These coefficients suggest that the Ridge learner uses the predictions of the four models in a corrective combination, rather than treating each model as equally useful.

5.3 Cross-Validation Analysis

The five validation folds produce R^2 values of 0.778, 0.779, 0.783, 0.771, and 0.765. The narrow spread is reflected in the mean ± standard deviation of 0.775 ± 0.007. The corresponding RMSE values range from 1.00 to 1.03 kcal/mol. This consistency supports the view that the reported performance is not dependent on a single favorable split.

5.4 Ablation Study

An ablation analysis was used to examine the effect of removing major components of the proposed pipeline. The strongest degradation is associated with removing target normalization, which returns the R^2 to approximately the baseline value. Removing target encoding reduces R^2 to 0.72, while replacing stacking with simple averaging produces $R^2 = 0.657$. Using only ECFP4 rather than the richer feature set gives $R^2 = 0.68$.

Table 7: Ablation Study

Configuration	R^2	Change Relative to Full Model
Full proposed model	0.794	—
Without target normalization	0.490	−0.304
Without target encoding	0.720	−0.074
Without stacking; simple average	0.657	−0.137
Without rich features; ECFP4 only	0.680	−0.114

The ablation results indicate that target-aware normalization is particularly important for this multi-target dataset. The feature representation and the learned stacking stage also make measurable contributions.

5.5 Comparison to Related Approaches

The reported result is competitive with several published binding-affinity prediction studies, although direct numerical comparison should be made cautiously because datasets, evaluation protocols, and input representations differ. Pafnucy reported $R^2 = 0.82$ on PDBbind, DeepDTA reported $R^2 = 0.64$ on KIBA, and OnionNet reported $R^2 = 0.71$ on BindingDB. The present system obtains $R^2 = 0.794$ on BindingDB using traditional machine-learning models rather than a deep neural architecture.

5.6 Computational Efficiency

The reported training workflow is comparatively lightweight. Feature generation requires approximately five minutes on the full dataset, while each base-model training stage takes about 90 seconds. Ridge meta-learner training takes less than one second, resulting in an overall training time of approximately ten minutes in the reported environment. Once trained, the prediction time is below ten milliseconds per molecule. These characteristics make the approach suitable for exploratory high-throughput screening workflows.

5.7 Limitations

- BindingDB measurements contain experimental noise and may be unevenly distributed across protein families, which can introduce dataset bias.
- Three-dimensional molecular information is not included because efficient 3D conformation generation was outside the reported implementation scope.
- CatBoost was not incorporated because of the stated environment constraints.
- Neural-network integration was not completed because of reported TensorFlow compatibility issues.
- The uncertainty transformation is a practical heuristic and should not be interpreted as a fully calibrated probabilistic confidence estimate.

5.8 Future Work

- Introduce efficient three-dimensional molecular representations and evaluate their incremental value.
- Add CatBoost as another diverse base learner.
- Perform systematic hyperparameter optimization, for example with Optuna.
- Develop target-specific models for proteins with sufficient training observations.
- Investigate graph neural networks and other learned molecular representations.

6. Conclusion

CAMELEON-AI X v3.0 provides a target-aware stacking framework for protein-ligand binding affinity

prediction. The method combines rich molecular fingerprints and descriptors with target statistics, target-stratified normalization, four diverse tree-based regressors, and a Ridge meta-learner trained from out-of-fold predictions. On the reported BindingDB test set, the approach reaches $R^2 = 0.7942$, RMSE = 0.9712 kcal/mol, and MAE = 0.7058 kcal/mol. The result represents a 62.1% relative R^2 improvement over the RandomForest baseline and a clear improvement over simple averaging.

The cross-validation mean of $R^2 = 0.775 \pm 0.007$ further indicates stable performance across data partitions. The analysis also shows that target-aware normalization, richer molecular features, and learned stacking each contribute to the final performance. The framework therefore offers a practical balance between predictive performance, interpretability of the ensemble coefficients, and computational cost.

11. References

1. DiMasi J.A., Grabowski H.G., Hansen R.W., “Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs”, Journal of Health Economics, 2016, 47, 20–33.
2. Kitchen D.B., Decornez H., Furr J.R., Bajorath J., “Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications”, Nature Reviews Drug Discovery, 2004, 3 (2), 154–166.
3. Ladbury J.E., Klebe G., Freire E., “Adding Calorimetric Data to Decision Making in Lead Discovery: A Hot Tip”, Nature Reviews Drug Discovery, 2010, 9 (1), 23–27.
4. Walters W.P., Barzilay R., “Applications of Deep Learning in Molecule Generation and Molecular Property Prediction”, Accounts of Chemical Research, 2021, 54 (2), 263–270.
5. Gaulton A., et al., “The ChEMBL Database in 2017: An Integrated Resource for Medicinal Chemistry and Drug Discovery”, Nucleic Acids Research, 2017, 45 (D1), D945–D954.
6. Wolpert D.H., “Stacked Generalization”, Neural Networks, 1992, 5 (2), 241–259.
7. Cheng T., et al., “BindingDB: A Public Database for Measuring Affinities of Protein-Ligand Complexes”, Nucleic Acids Research, 2018, 46 (D1), D564–D569.
8. Rogers D., Hahn M., “Extended-Connectivity Fingerprints”, Journal of Chemical Information and Modeling, 2010, 50 (5), 742–754.
9. Sheridan R.P., “Time-Split Cross-Validation as a Method for Estimating the Uncertainty of Performance Computed from Random Data Splitting”, Journal of Chemical Information and Modeling, 2019, 59 (10), 4436–4452.
10. Breiman L., “Stacked Regressions”, Machine Learning, 1996, 24, 49–64.
11. Tropsha A., “Best Practices in QSAR Model Development, Validation and Exploitation”, Molecular Informatics, 2010, 29 (6–7), 476–488.
12. Stepniewska-Dziubinska M.M., Zielenkiewicz P., Siedlecki P., “Development and Validation of the Deep Learning Model for the Prediction of Protein-Ligand Binding Affinity”, Journal of Chemical Information and Modeling, 2018, 58 (12), 2441–2452.
13. Öztürk H., Özgür A., Ozkirimli E., “DeepDTA: Deep Drug-Target Affinity Prediction”, Bioinformatics, 2018, 34 (17), i821–i829.
14. Liu Y., et al., “OnionNet: A Deep Multiple-Layer Learning Method for Predicting Drug-Target Binding Affinity via 3D Structure”, Bioinformatics, 2019, 35 (24), 5148–5156.