BioActML: Predicting mTOR Inhibitor Bioactivity with Machine Learning

Abstract

This report details the development and evaluation of a machine learning pipeline, BioActML, designed to predict the bioactivity (pIC_{50}) of chemical compounds against the protein target mTOR. Data was sourced from the ChEMBL database, and compounds were represented numerically using 1024-bit Morgan fingerprints. Two regression models, Random Forest and Gradient Boosting, were trained and evaluated using a 5-fold cross-validation strategy. The Random Forest model emerged as the superior predictor, achieving a coefficient of determination (R^2) of 0.226 and a Root Mean Squared Error (RMSE) of 1.019. Further analysis through visualization of the chemical space with PCA and t-SNE confirmed that the chosen features captured meaningful structure-activity relationships. This work establishes a robust baseline for predicting mTOR inhibitor potency from chemical structure alone.

---

1. Introduction

The identification of small molecules that modulate the activity of protein targets is a cornerstone of modern drug discovery. The Mammalian Target of Rapamycin (mTOR) is a critical kinase involved in cell growth and proliferation, making it a key target for therapeutic development, particularly in oncology. Quantifying the potency of compounds against a target, often expressed as the pIC_{50} value, is essential for prioritizing candidates. This project, BioActML, aims to construct a machine learning workflow to predict the pIC_{50} of compounds targeting mTOR, leveraging publicly available data from the ChEMBL repository. The goal is to build and validate a model that can accurately predict bioactivity directly from a compound's 2D chemical structure.

---

2. Methods

2.1. Data Acquisition and Preprocessing

Bioactivity data was collected from the ChEMBL database, an open-access resource of drug-like bioactive molecules. The specific protein target selected was **mTOR (ChEMBL ID: CHEMBL2842).** All compound-target binding data for human targets with valid IC_{50} measurements were downloaded using the ChEMBL Web Resource Client API. The dataset was filtered to retain only entries with canonical SMILES strings. The IC_{50} values, provided in nanomolar (nM) units, were converted to the logarithmic pIC_{50} scale using the transformation $pIC_{50} = -\log_{10}(IC_{50}[M])$. The final cleaned dataset of compound SMILES and their corresponding pIC_{50} values was saved to a CSV file for further processing.

2.2. Featurization

To make the dataset compatible with machine learning models, the SMILES strings were converted into numerical descriptors. This was accomplished using the RDKit library to generate Morgan fingerprints, a type of circular fingerprint. Each molecule was encoded into a 1024-bit binary vector where each bit corresponds to the presence or absence of a specific chemical substructure within a radius of 2. Molecules for which fingerprint generation failed were excluded. This procedure resulted in a feature matrix X of shape (n_samples, 1024) and a target vector y containing the pIC_{50} values.

2.3. Model Training and Evaluation

The core objective was to develop a supervised regression model capable of predicting compound bioactivity (pIC_{50}) from the molecular fingerprints generated previously.

To ensure a robust assessment of model generalizability, a 5-fold cross-validation strategy was employed. This technique provides more reliable performance estimates than a single train-test split. Model performance was quantified using three standard metrics:

- **R^2 (coefficient of determination):** Measures the proportion of variance in the target variable that is predictable from the features.
- **RMSE (Root Mean Squared Error):** A measure of prediction accuracy that heavily penalizes larger errors.
- **MAE (Mean Absolute Error):** The average absolute difference between predicted and true values.

Two models were evaluated:

1. **Random Forest Regressor:** This model, configured with 200 trees, served as the baseline. It is known for its strong performance and resistance to overfitting on tabular data.
2. **HistGradient Boosting Regressor:** This is a gradient boosting method optimized for large datasets. Unlike Random Forest, it builds trees sequentially, allowing each tree to correct the errors of its predecessors. The model was first tested with default hyperparameters and then tuned using GridSearchCV across a parameter grid for max_iter, learning_rate, and max_leaf_nodes.

---

3. Results and Discussion

3.1. Dataset Characteristics

An initial analysis of the target variable, pIC_{50}, is crucial for understanding the data distribution the model will learn from. The distribution of pIC_{50} values in the final dataset is shown in Figure 1.

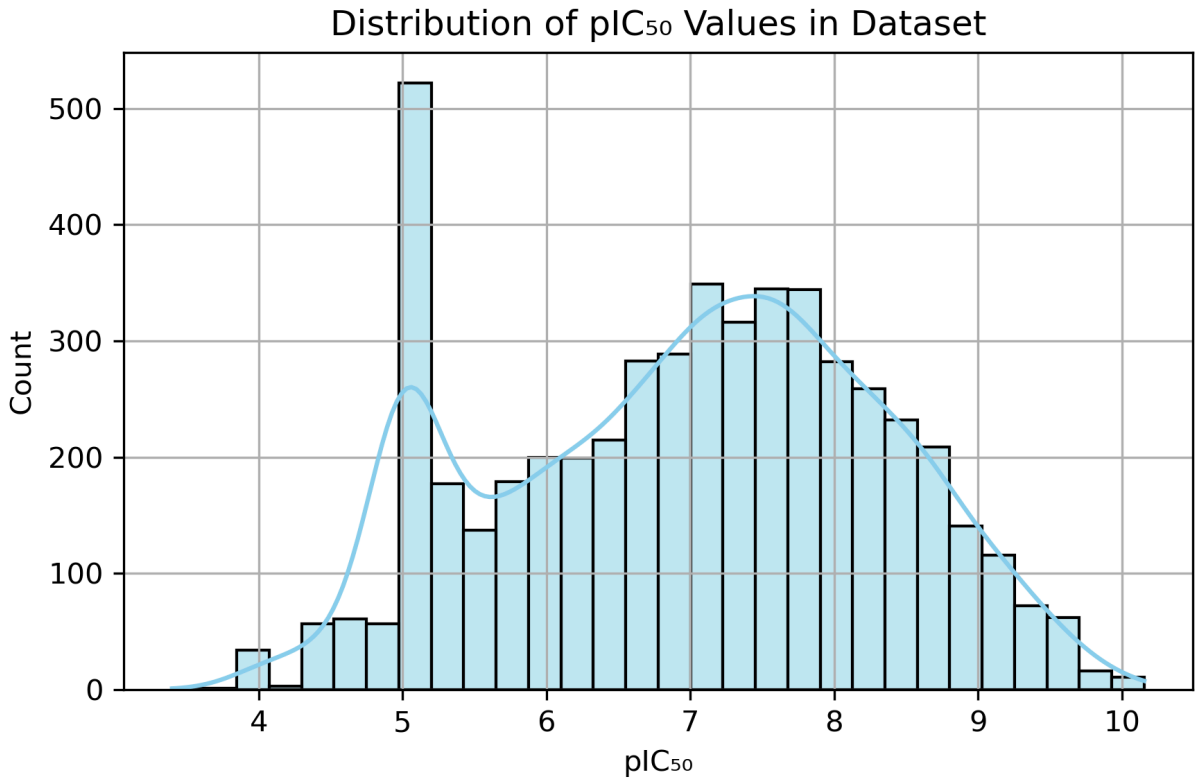


Figure 1: Distribution of pIC_50 Values in the Dataset.

The distribution is notably skewed, featuring a sharp peak around a pIC_{50} value of 5 and a second, broader distribution across the 6-9 range. This indicates that the dataset contains a mix of compounds with low, moderate, and high activity, but with a significant cluster of compounds near the threshold for weaker binding affinity. This dense peak could reflect biases from data reporting or specific assays and may influence model training by overrepresenting low-potency examples.

3.2. Model Performance Comparison

The performance of the Random Forest, default Gradient Boosting, and tuned Gradient Boosting models was evaluated using 5-fold cross-validation. The results are summarized in Table 1.

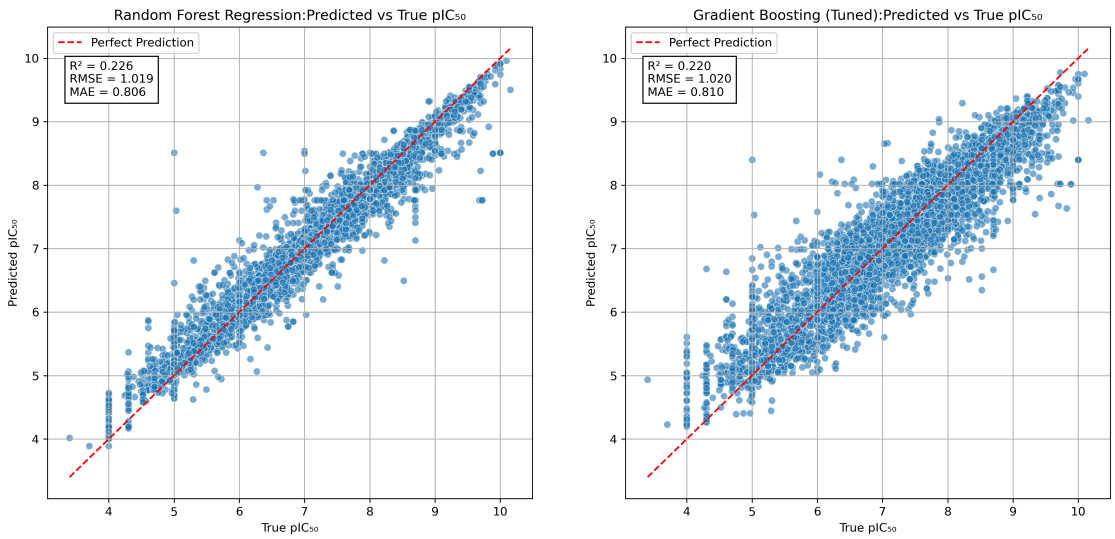
Table 1: Model Performance Comparison

Model	R2	RMSE	MAE
Random Forest (Baseline)	0.226	1.019	0.806

Gradient Boosting (Default)	0.190	1.038	0.812
Gradient Boosting (Tuned)	0.220	1.020	0.810

The Random Forest Regressor achieved the highest overall performance, with an R^2 of 0.226, an RMSE of 1.019, and an MAE of 0.806. This indicates the model explains approximately 22.6% of the variance in pIC_{50} values. The default Gradient Boosting model underperformed relative to the Random Forest model. Although hyperparameter tuning provided a noticeable improvement to the Gradient Boosting model, it did not surpass the Random Forest baseline. This suggests that the ensemble averaging approach of Random Forest may be better suited for capturing the structure-activity relationships in this particular dataset and feature representation.

The relationship between predicted and true pIC_{50} values for the two best models is visualized in Figure 2.



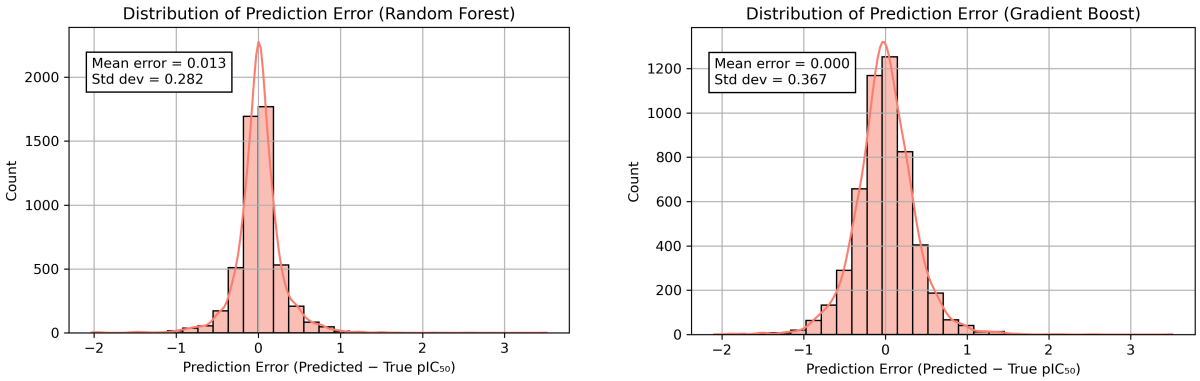
(a) Random Forest (b) Tuned Gradient Boosting

Figure 2: pIC_{50} Metrics shown are from 5-fold cross-validation.

Visually, both models show a dense clustering of predictions around the diagonal, confirming they have captured meaningful signals from the data. The slightly better performance of the Random Forest is visible in the tighter grouping of points around the line of perfect prediction.

3.3. Prediction Error Analysis

To better understand model bias, the distribution of prediction errors (Predicted - True pIC_{50}) was analyzed (Figure 3).



(a) Random Forest (b) Tuned Gradient Boosting

Figure 3: Distribution of Prediction Errors

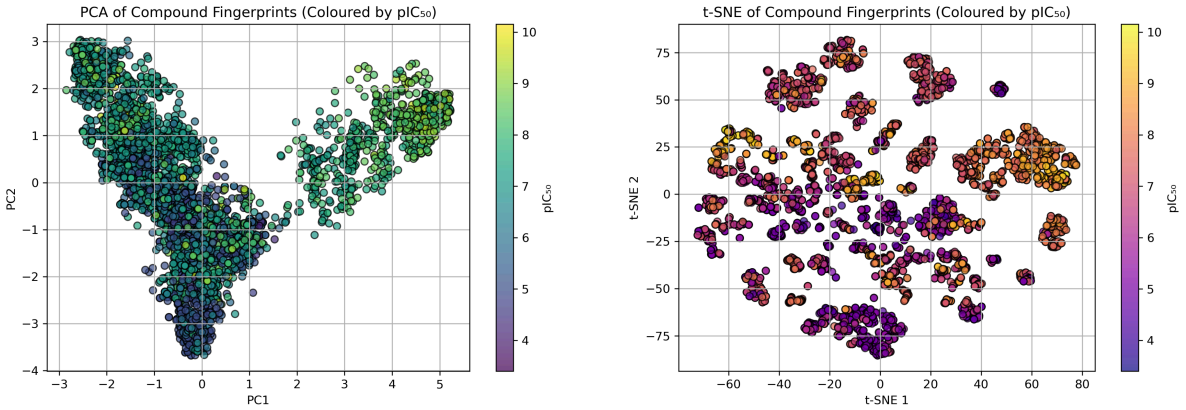
Both error distributions are unimodal and approximately symmetric, suggesting neither model has a strong systematic bias toward over- or under-prediction. The Random Forest model displays a tighter error distribution with a lower standard deviation (0.282 vs. 0.367 for Gradient Boosting), indicating its predictions are more consistently close to the true values. Conversely, the Gradient Boosting model is slightly better calibrated, with a mean error of 0.000 compared to the Random Forest's 0.013.

3.4. Chemical Space Visualization

To investigate the structural diversity of the dataset, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) were used to project the 1024-dimensional Morgan fingerprints into two dimensions.

The PCA plot (Figure 4a) provides a view of the global variance in the dataset's structural features. The plot reveals distinct structural clusters, suggesting the fingerprints capture chemically meaningful similarities. Importantly, smooth color gradients are visible in several regions, indicating a continuous structure-activity relationship where small structural changes lead to predictable changes in bioactivity.

The t-SNE plot (Figure 4b) is a nonlinear technique that excels at preserving local similarities between compounds. It reveals a rich landscape of tightly-defined clusters that are more visually separated than in the PCA plot. Many of these clusters are "activity-coherent," meaning compounds within them share similar potency levels, which reinforces that the local chemical environments captured by the fingerprints are highly relevant for predicting bioactivity.



(a) PCA of Compound Fingerprints (b) t-SNE of Compound Fingerprints

Figure 4: PCA and t-SNE visualisation

Together, these visualizations confirm that the Morgan fingerprint feature space is not random but encodes chemically and biologically meaningful relationships that align with bioactivity, justifying the modeling strategy.

3.5. Feature Importance Analysis

To interpret which structural features the best model (Random Forest) learned to prioritize, the importance of each of the 1024 bits in the Morgan fingerprint was analyzed. Figure 5 shows the top 20 most important fingerprint bits for the model's predictions.

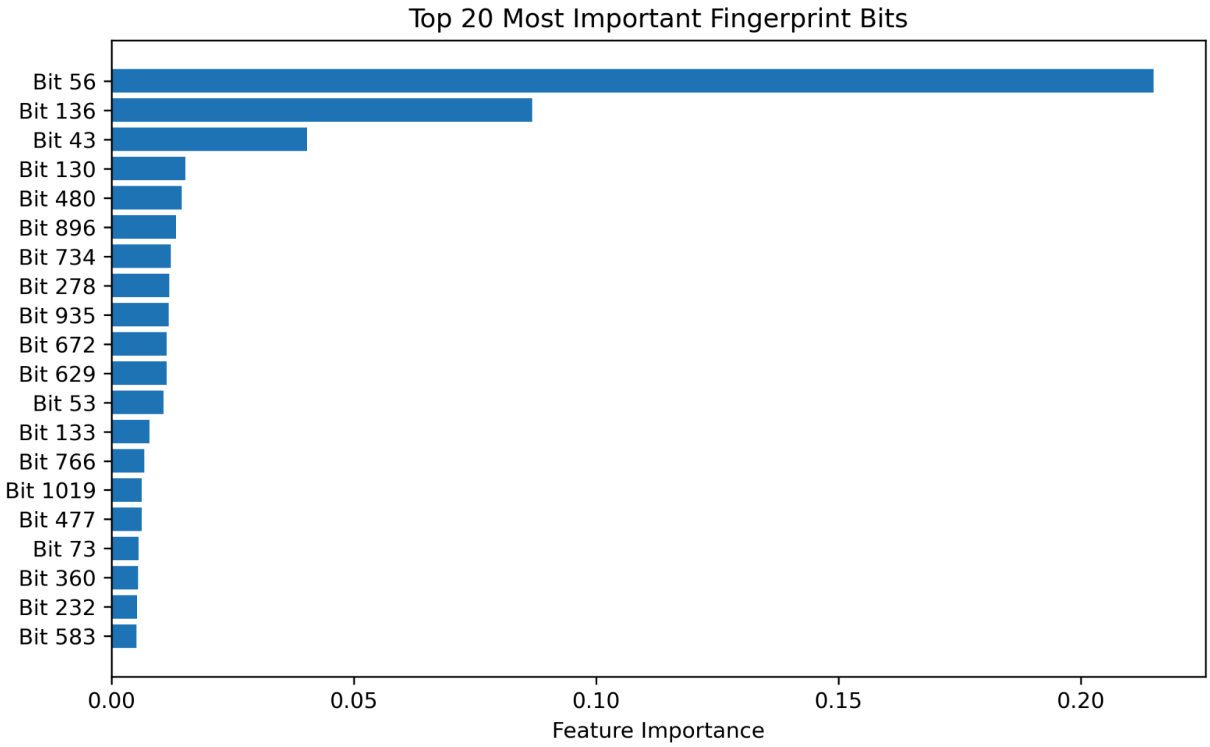


Figure 5: Top 20 most important Morgan fingerprint bits as determined by the Random Forest model.

This analysis reveals that a small subset of features has a disproportionately large influence on the model's predictions. For example, Bit 56 is significantly more important than any other feature. In a real-world scenario, these bits could be mapped back to the specific chemical substructures they represent, providing valuable insights for medicinal chemists to guide new compound design.

---

4. Conclusion

This project successfully developed and evaluated a machine learning pipeline, BioActML, to predict the bioactivity (pIC_{50}) of compounds against the mTOR protein target. A Random Forest model trained on Morgan fingerprints demonstrated the best predictive performance, achieving an R2 of 0.226. While this level of performance indicates the model has captured a significant signal, it also highlights the inherent complexity of predicting bioactivity from 2D structure alone. Visualizations of the chemical space via PCA and t-SNE confirmed that the featurization strategy captured meaningful structure-activity relationships.

Future work could focus on improving predictive accuracy by incorporating more sophisticated features, such as physicochemical descriptors or 3D structural information. Additionally, more advanced modeling techniques, such as model stacking or graph neural networks, could be explored to potentially capture more complex relationships within the data.