Multi-Modal Modeling of Compound Bioactivity: A Comparative Study of Regression and Classification Approaches

Group 4

Dennis Mathew Jose
Venkata Siva Naga Vamsinath Thatha

jose.de@northeastern.edu
thatha.v@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Dennis Mathew Jose

Signature of Student 2: Venkata Siva Naga Vamsinath Thatha

Submission Date: 04/22/2025

Abstract

Drug discovery remains a vital yet highly resource-intensive endeavor, traditionally dependent on the experimental screening of large libraries of chemical compounds. This approach is often time-consuming, costly, and inefficient. To address these challenges, this project explores the use of artificial intelligence and machine learning to predict the bioactivity of chemical compounds, aiming to streamline the early stages of drug development. Leveraging the ChEMBL database, which contains extensive data on compound-target interactions, the study focuses on predicting pIC50 values—a widely used measure of a compound's inhibitory strength.

The initial phase of the project framed the problem as a regression task. Several models were trained and evaluated, including Ridge Regression, Random Forest, Multi-Layer Perceptron (MLP), and XGBoost. Incorporating Morgan fingerprints, which capture substructural information about molecules, led to notable improvements in model accuracy. Among the models tested, XGBoost demonstrated the most reliable performance. In parallel, SHAP (SHapley Additive exPlanations) was employed to enhance interpretability, allowing for a clear understanding of how individual features contributed to each prediction.

To align more closely with decision-making practices in drug screening, the regression task was subsequently reformulated as a binary classification problem. Compounds were labeled as active or inactive based on a biologically relevant pIC50 threshold. Classification models included Logistic Regression as a baseline, along with Random Forest, XGBoost, and MLP classifiers. This dual approach enabled both continuous and categorical predictions, providing flexibility for various stages of the drug discovery pipeline.

This project has the potential to accelerate the development of new therapies for diseases like cancer, Alzheimer's, and diabetes. By combining cutting-edge AI techniques with large-scale biological data, we can make drug discovery more efficient, affordable, and accessible, ultimately improving global health outcomes. The model will benefit pharmaceutical companies**,** academic researchers**,** healthcare providers, and patients by reducing R&D costs, shortening development timelines, and improving the success rate of drug candidates in clinical trials.

1. Project Definition

Drug discovery is essential for developing new medicines to treat various diseases, but the process is often slow, costly, and inefficient. Traditional methods rely on high-throughput screening (HTS) and laboratory experiments to test thousands of chemical compounds, requiring years of research and billions of dollars in investment. Many drug candidates fail during clinical trials due to toxicity, poor efficacy, or unforeseen side effects, leading to significant financial losses and delays in bringing effective treatments to patients. This inefficiency highlights the need for more advanced approaches that can streamline early-stage drug screening and improve success rates.

Artificial intelligence (AI) and machine learning (ML) have emerged as powerful tools to enhance drug discovery by predicting how well a chemical compound interacts with a biological target before experimental testing. By leveraging large-scale datasets, ML models can analyze

molecular structures, bioactivity data, and physicochemical properties to identify promising drug candidates more efficiently. Cheminformatics, which combines computational techniques with chemical data, plays a crucial role in this approach by providing systematic methods to evaluate molecular descriptors that influence drug-target interactions.

The increasing availability of chemical and biological data, such as that found in the ChEMBL database, enables researchers to develop predictive models that estimate drug bioactivity more accurately. By using AI-driven predictions, early-stage drug screening can become more targeted, reducing the number of ineffective compounds undergoing costly experimental validation. This not only accelerates drug development but also helps optimize the selection of drug-like compounds with favourable properties such as solubility and minimal toxicity. Beyond efficiency, AI-driven drug discovery has significant implications for personalized medicine, where treatments can be tailored based on individual genetic profiles.

This project aims to build a machine-learning model to predict drug bioactivity (pIC50) using molecular descriptors and historical bioactivity data from the ChEMBL database. By integrating AI, cheminformatics, and bioinformatics techniques, this study seeks to enhance early-stage drug discovery, reduce costs, and increase the likelihood of successful drug candidates progressing to clinical trials. This approach has the potential to transform pharmaceutical research by improving the efficiency and affordability of drug development, ultimately benefiting healthcare providers and patients worldwide.

2. Problem Statement

Accurately predicting compound bioactivity is a fundamental challenge in drug discovery, as it directly influences the selection of potential therapeutic candidates. Experimentally, compound bioactivity is measured using IC50 assays, which quantify the concentration needed to inhibit 50% of a target protein's activity. These assays use a range of techniques, including fluorescence-based detection, radioligand binding, and enzymatic inhibition. To standardize outputs across studies, the pIC50 metric—the negative logarithm (base 10) of the IC50 value—is used, with higher pIC50 values indicating greater potency. This project addresses the specific problem of predicting compound bioactivity against target proteins using data from the ChEMBL database. The goal is to build a machine learning model that can accurately predict compounds' pIC50 (negative logarithm of the IC50 value), which measures their binding affinity for target proteins and then classify them as active or inactive compounds.

This project addresses the problem of predicting bioactivity by building machine learning models that estimate pIC50 values using molecular descriptors and compound-target data sourced from the ChEMBL database. While these features capture general chemical properties, they may not fully represent the complex substructural patterns that drive biological activity. To enhance model performance beyond the baseline, Morgan fingerprints were introduced. These circular fingerprints encode detailed information about a molecule's substructures and connectivity, helping the model better capture structure-activity relationships. These regression models aim to provide accurate, continuous estimates of inhibitory potency, supporting the early prioritization of compounds before experimental screening. However, real-world drug discovery workflows often involve binary decision-making, such as whether a compound should proceed to the next stage of

validation. To align the modeling approach with this practical need, the project also reframes the task as a classification problem, labeling compounds as active or inactive based on a biologically meaningful pIC50 threshold.

The central research questions addressed in this project are:

1. Can machine learning models accurately predict the bioactivity (pIC50) of compounds using molecular descriptors and target-related information?
2. How effectively can binary classification models distinguish between active and inactive compounds?
3. What molecular features—including both general physicochemical descriptors and structural fingerprints—are most predictive of bioactivity?

Addressing this issue is as crucial as it can lower the time and expenses involved in drug discovery by allowing researchers to concentrate on the most promising compounds. Furthermore, it offers insights into the molecular mechanisms behind drug-target interactions, fostering the creation of more effective and safer medications. The potential impact of this project reaches the pharmaceutical industry, healthcare providers, and patients, as it can hasten the introduction of new therapies to the market.

3. Overview of the Data and Structure

3.1 Data Source

The primary data source for this project is the ChEMBL database, a publicly available resource containing bioactivity data for over 2 million compounds and 15,000 targets. The data includes:

- Compound information: Molecular structures (SMILES format), molecular weight, and lipophilicity (logP).
- Bioactivity data: IC50, Ki, and other bioactivity measurements.
- Target information: Protein targets, organisms, and assay details.

The ChEMBL database is accessed via its web interface and API, which allow users to query and download data programmatically. The data is well-documented and widely used in the cheminformatics and drug discovery communities. It is considered a credible and reliable source due to its comprehensive coverage, regular updates, and extensive use in academic and industrial research.

3.2 Data Description

Dataset Overview

- **Number of Rows**: 10,000 (initially, can be expanded).
- **Number of Columns**: 15-20 (depending on feature engineering).

The dataset used in this project is derived from the ChEMBL database and includes the following variables:

molecule_chembl_id	Categorical	Unique identifier for the compound.
target_chembl_id	Categorical	Unique identifier for the target protein.
canonical_smiles	Text	SMILES representation of the molecular structure.
standard_value	Numeric	IC50 value (measure of bioactivity in nM).
pIC50	Numeric	Negative logarithm of the IC50 value (target variable).
assay_chembl_id	Categorical	Identifier for the assay used to measure bioactivity.
organism	Categorical	Organism the target protein belongs to (e.g., human, mouse).
target_family	Categorical	Family of the target protein (e.g., kinase, GPCR, ion channel).
assay_type	Categorical	Type of assay used (e.g., binding, functional).
compound_class	Categorical	Class of the compound (e.g., small molecule, peptide, natural product).
num_h_donors	Numeric	Number of hydrogen bond donors in the compound.
num_h_acceptors	Numeric	Number of hydrogen bond acceptors in the compound.
tpsa	Numeric	Topological polar surface area (in Å²).
num_rotatable_bonds	Numeric	Number of rotatable bonds in the compound.
lipinski_compliant	Binary	Indicates whether the compound complies with *Lipinski's Rule of Five (1 = yes, 0 = no).
toxicity_risk	Ordinal	Toxicity risk level of the compound (1 = low, 5 = high).
solubility_level	Ordinal	Solubility level of the compound (1 = low, 5 = high).

Table 3.1: Describe the Variables, data type and descriptions

(Here we have used 17 variables for feature engineering. The number of feature variables can be increased if we require a more detailed analysis which is explained in the subsequent sections. Also added 512-bit Morgan fingerprints after baseline modeling to enhance the predictive power of the model.)

The molecular descriptors and bioactivity measurements are not obtained at the same time. Molecular descriptors (e.g.,TPSA, hydrogen bond donors/acceptors) are computed from the chemical structure of the compound using cheminformatics tools. These values are fixed once the structure of the molecule is known and do not require experimental testing. Bioactivity measurements (e.g., IC50, pIC50) are experimentally determined in the lab using biochemical or cellular assays. These measurements depend on how the compound interacts with a specific biological target under experimental conditions. The data of molecular descriptors and bioactivity values are later integrated for machine learning analysis.

3.3 Data Structure

The dataset is structured as a tabular format with rows representing individual compound-target interactions and columns representing features and target variables. The target variable is pIC50, which is derived from the standard_value (IC50) using the formula:

$$\boldsymbol{pIC50} = -log_{10}(\boldsymbol{IC50})$$

The dataset includes a mix of numeric, categorical, and ordinal features, making it suitable for advanced machine-learning models.

4. Data Visualization and Pre-Processing

4.1 Data Exploration

The dataset includes various features, encompassing both numerical and categorical types. The target variable, pIC50, represents the potency of the compounds, and our goal is to build a model to predict this value.

4.1.1 Missing Data Overview

Upon initial inspection, it was found that some features had missing values. The visualization of missing value helps us to understand which columns have missing data and their extent. Columns with high proportions of missing data may require imputation or removal.

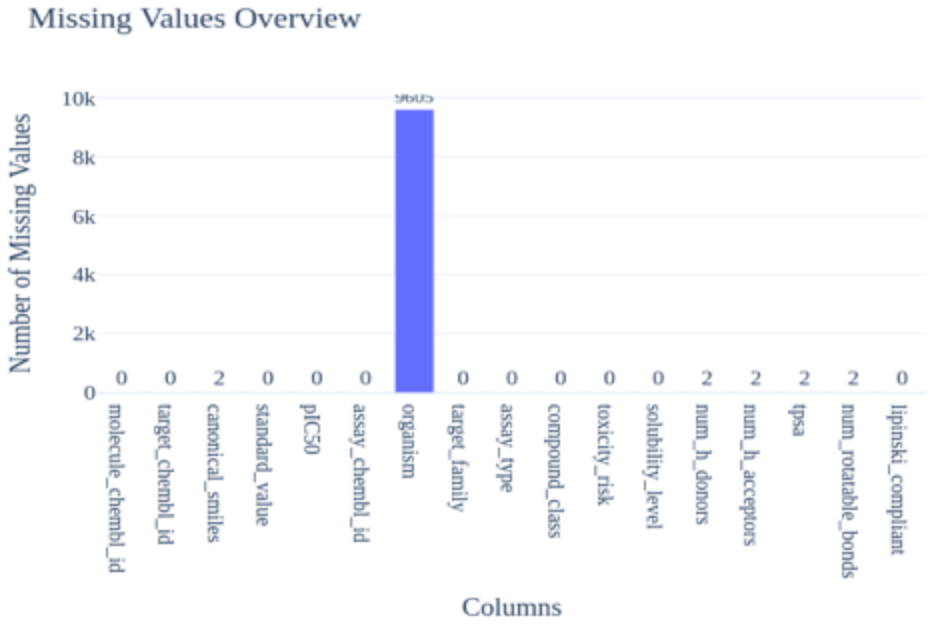

fig 4.1: Missing values overview

Upon reviewing the dataset, we found that the Organism column contains 100% missing values and does not contribute any valuable information for the analysis. Therefore, we have decided to drop this column to streamline the dataset and avoid complications.

The remaining columns' missing values are minimal (0.02%), which is unlikely to impact the analysis significantly. The SMILES column contains crucial information for molecular representation. Since this column is essential for further analysis, rows with missing SMILES values have been dropped entirely. Dropping these rows ensures that the dataset remains relevant and focused on the most accurate molecular representations.

After handling for these two columns if there are any missing values, we have chosen to impute the missing values to preserve data integrity and ensure that the dataset remains complete for analysis. The choice of imputation method depends on the type of data in each column.

Numerical missing values have been imputed using the median, which is robust to outliers and ensures that the imputed values are representative of the data distribution without introducing bias. Categorical missing values have been imputed with the mode (the most frequent category), which is a common and effective approach for handling categorical data.

4.1.2 Distribution of pIC50 Values and Dealing with outliers

The distribution of pIC50 values appears to be primarily right skewed, with most values concentrated between approximately 3 and 10. The histogram shows a bell-shaped distribution with a peak around 5-6, but there's also a small number of negative values (between -10 and 0) that appear as outliers.

The boxplot at the top confirms this distribution pattern, with:

- The interquartile range (the blue box) sitting between approximately 4 and 7
- The median (vertical line in the box) around 5-6
- A few outliers visible on the left side (negative values)

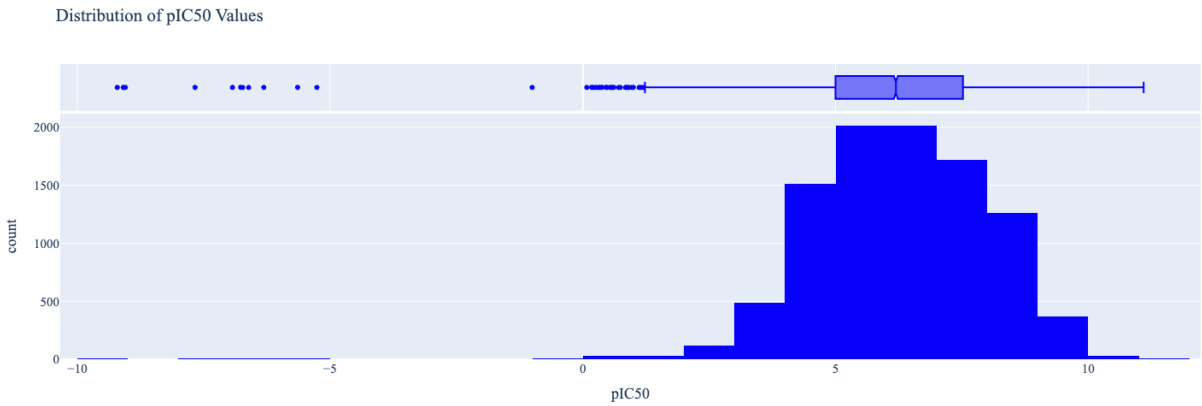

Fig 4.2: Distribution of pIC50 values and outlier detection of pIC50

pIC50 is typically calculated as -log10(IC50), where IC50 is measured in molar concentration. Negative pIC50 values would correspond to IC50 values greater than 10M, which is extremely high and often biologically implausible or meaningless for most drug discovery contexts. The main distribution (positive values) likely represents compounds with meaningful biological activity, while the negative outliers might represent inactive compounds or experimental artifacts.

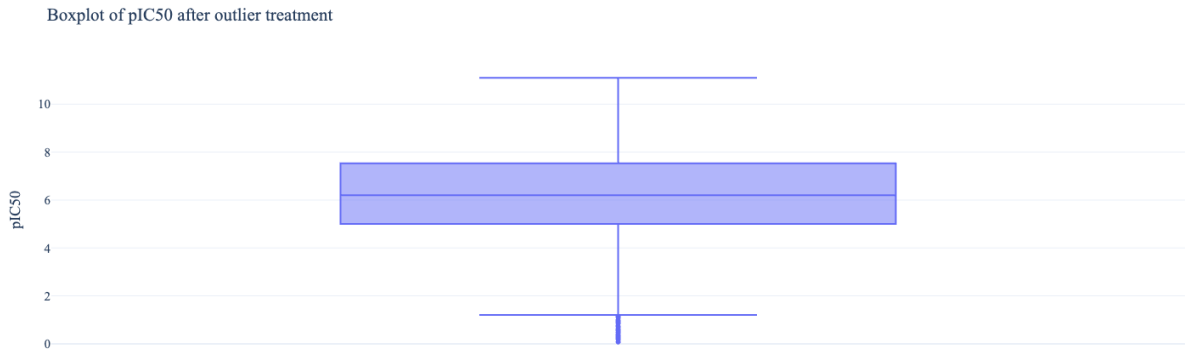

Fig 4.4: Box plot for pIC50, after treating the negative outliers

Molecular Property Outliers:

- Hydrogen Bond Features: Both num_h_donors and num_h_acceptors show strong positive skew with most compounds having low values (0-5), but with notable outliers having values as high as 40-60.
- TPSA (Topological Polar Surface Area): Shows similar skewed distribution with extreme outliers reaching 1500+, while most compounds cluster below 200.

- Rotatable Bonds: Most compounds have fewer than 10 rotatable bonds, but outliers exist with 80+ rotatable bonds.

Toxicity and Solubility: These appear more normally distributed with values typically between 1-5 and fewer extreme outliers compared to molecular descriptors.

Lipinski Compliance: This appears to be a binary or categorical feature (likely 0/1), indicating whether compounds follow Lipinski's rule of five*.

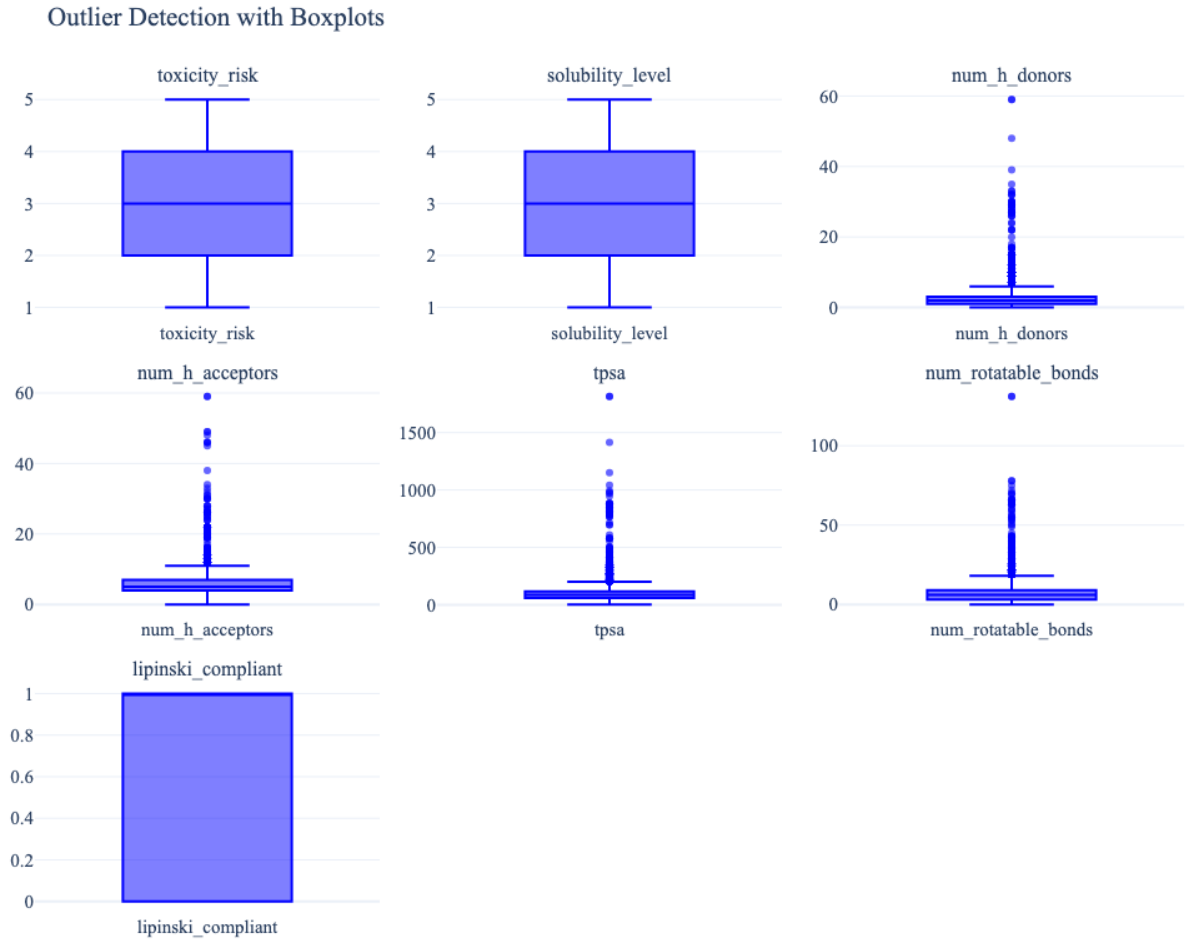

Fig 4.3: Outlier Detection of Numerical Features

The presence of these outliers in the ChEMBL dataset suggests they passed some level of quality control and might represent genuine chemical space that's underexplored in drug discovery. These outliers should be examined in context rather than automatically removed, as they might represent valuable chemical matter with unique properties or activity profiles.

5. Feature Engineering

Feature engineering plays a critical role in our drug discovery project by transforming raw molecular descriptors into more meaningful representations that better capture structure-

activity relationships. Given the complex nature of ChEMBL bioactivity data and the skewed distributions observed in molecular properties such as hydrogen bond donors/acceptors, TPSA, and rotatable bonds, we applied various feature engineering techniques to optimise our dataset for modelling.

5.1 Encoding of Categorical Variables

For encoding categorical data, we used the LabelEncoder from sklearn.preprocessing to convert categorical variables into numerical representation. This is necessary as machine learning algorithms typically require numerical input. The columns 'target_chembl_id', 'compound_class',' target_family' and 'assay_type' were encoded by applying the encoder to each column, transforming their categorical values into integer labels that retain the information while making the dataset suitable for further analysis.

5.2 Scaling and standardizing the data

In the case of scaling and standardizing the numerical data, we utilized the StandardScaler from sklearn.preprocessing. This method ensures that the numerical features 'standard_value', 'num_h_donors', 'num_h_acceptors', 'tpsa', and 'num_rotatable_bonds' are standardized to have a mean of 0 and a standard deviation of 1. Standardization is crucial to ensure that all features contribute equally to the model, especially when they have different units or scales, preventing features with larger ranges from disproportionately influencing the model.

5.3 Correlation Analysis

The correlation matrix was calculated to understand the relationships between the numerical features. Strong correlations between features suggest potential multicollinearity, which could lead to redundant information in the model.

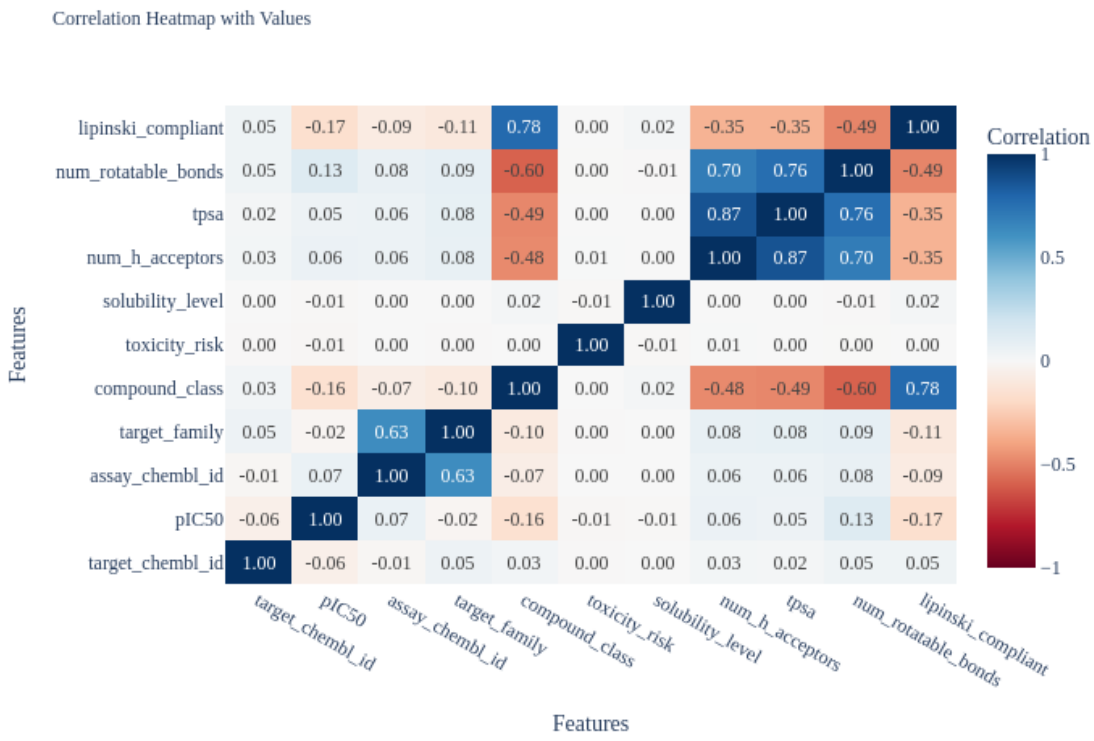

Fig 5.1: Correlation Matrix of features

Highly correlated features can lead to multicollinearity in regression models, which can adversely affect the performance and interpretability of the model. For example, the correlation of 0.95 between 'tpsa' and 'num_h_donors' indicates a strong linear relationship, which could cause issues like inflated standard errors and unreliable coefficient estimates.

5.4 Variance Inflation Factor over PCA

We opted to use the Variance Inflation Factor (VIF) rather than Principal Component Analysis (PCA) for detecting and removing multicollinear features. VIF is a more direct and interpretable method for identifying multicollinearity. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other features. A high VIF indicates that a feature is highly correlated with others, suggesting that it could be redundant.

We chose VIF over PCA because VIF retains the original features in their meaningful forms, allowing us to directly identify and eliminate the problematic variables. PCA, on the other hand, transforms features into new dimensions (principal components), making it harder to interpret the relationship between original features and model outputs. Also, PCA transformation is generally applied if there are many features, here the original dimension is not very huge. Therefore, VIF provides a clearer and more transparent approach to handle multicollinearity in this case.

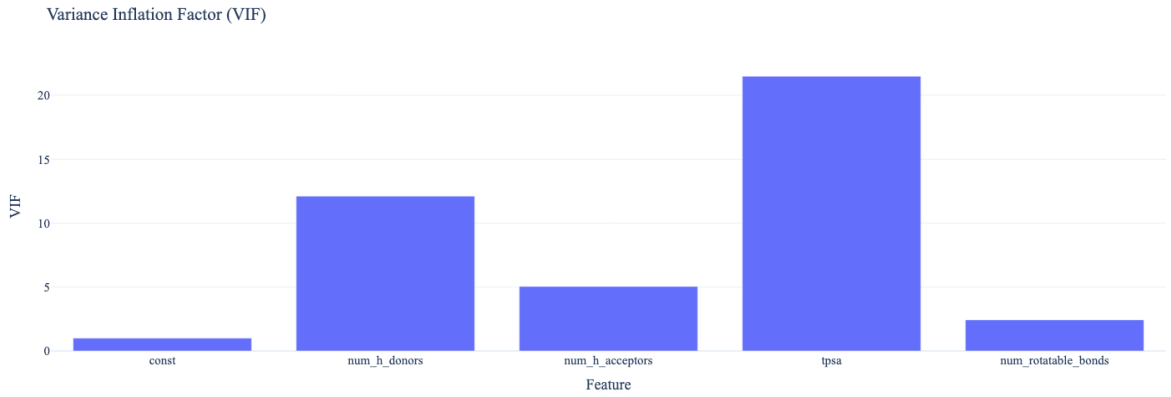

Fig 5.2: Multicollinearity Analysis using VIF

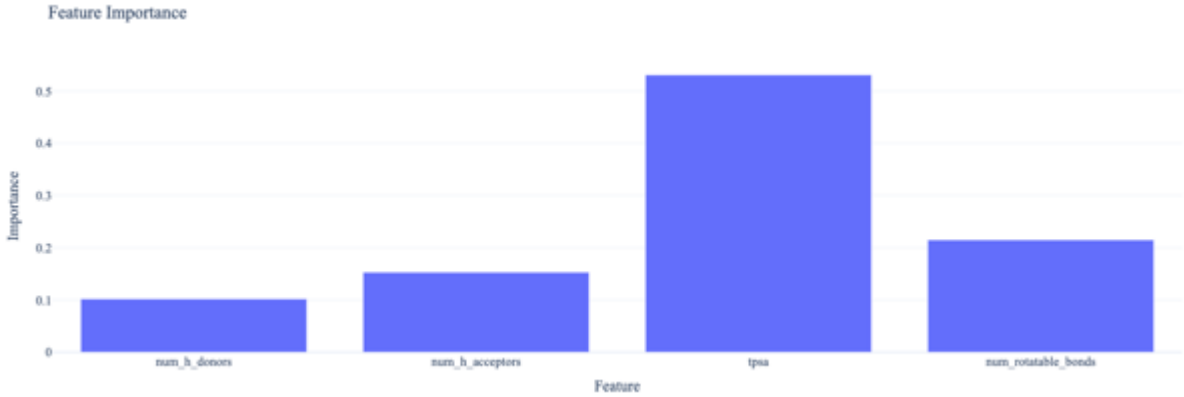

Fig 5.3: Feature Importance and Multicollinearity Analysis

Based on the analysis of the Variance Inflation Factor (VIF) and Feature Importance values, we can make informed decisions about which features to retain or remove from the model. The feature 'tpsa' has the highest VIF (21.51) and the highest feature importance (0.534), indicating that it is highly influential in the model and likely critical for retaining. Despite its high multicollinearity with 'num_h_donors', its importance suggests that it contributes significantly to the model's predictive power, so it should likely be kept.

On the other hand, 'num_h_donors' shows a high VIF (12.11) but has a relatively low feature importance (0.103). This indicates that although 'num_h_donors' is highly correlated with other features, it does not significantly contribute to the model's performance. Given its high VIF and low importance, it might be a candidate for removal, as its inclusion could add unnecessary complexity to the model without improving its predictive accuracy. Therefore, removing 'num_h_donors' would help reduce multicollinearity while maintaining the interpretability and performance of the model.

5.5 Engineering New Features

To enhance the predictive power of the model, new molecular features were generated based on existing SMILES (Simplified Molecular Input Line Entry System) representations of the compounds. By computing molecular descriptors, such as molecular weight (MolWt), lipophilicity (LogP), and the number of rings in the molecular structure (NumRings), additional information is

integrated into the dataset. Generating these new features is essential as they provide deeper insights into molecular properties that can significantly impact target interactions and assay outcomes, improving the model's predictive accuracy.

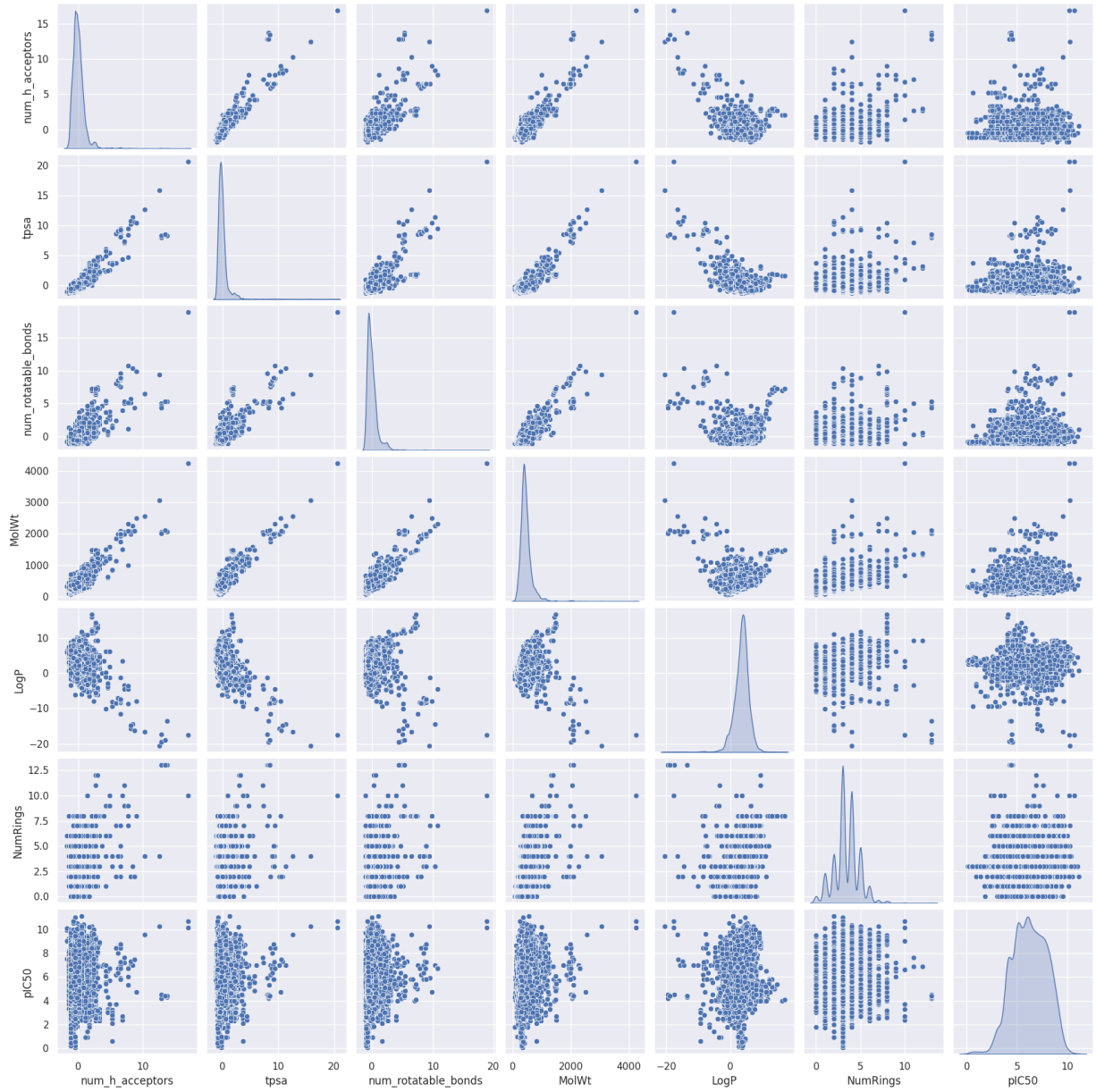

Fig 5.4 Pair plots

Upon examining the pair plot after addressing multicollinearity, several interesting patterns and trends emerge. The diagonal histograms indicate that descriptors like num_h_acceptors, tpsa, and num_rotatable_bonds exhibit right-skewed distributions, with most compounds having lower values. These descriptors are commonly associated with *drug-like properties, and the property space of the dataset largely falls within typical drug-like boundaries—molecular weight (MolWt) is below 500, LogP ranges between -5 and 5, and the number of rings and rotatable bonds is moderate. This suggests the dataset predominantly represents chemical space relevant to drug discovery.

Further analysis of the property relationships reveals that higher pIC50 values are slightly more common for compounds with moderate LogP values (0-5), although no strong linear correlations between individual descriptors and pIC50 exist. The patterns observed suggest that combinations of molecular properties, rather than any single property, are likely responsible for determining bioactivity. For instance, compounds with similar molecular weights can exhibit widely varying pIC50 values. Additionally, while correlations such as between num_h_acceptors and tpsa are chemically expected, the scatter patterns in the pairplot suggest that these features provide complementary rather than redundant information, contributing to a more nuanced understanding of bioactivity.

6. Modeling & Performance Evaluation

With the data pre-processed and descriptors generated, we transitioned into the modelling phase, where the objective was to predict compound bioactivity—first as a regression task to estimate pIC50, and subsequently reframed as a classification task (active vs. inactive). The modeling process aimed not only to achieve strong predictive performance but also to analyze, interpret, and compare the behaviour of various machine learning models under a real-world drug discovery context.

6.1 Data Splitting Strategy

To ensure robust and generalizable model performance, we partitioned the dataset into three distinct subsets:

- **Test Set (Hold-out)**: A final evaluation set of 1,000 records was isolated from the original dataset to assess model performance on unseen data.
- **Training and Validation Sets**: The remaining data was further divided in an 80:20 **ratio** using train_test_split, with 80% allocated for training the models and 20% reserved for validation.

This rigorous splitting strategy facilitated hyperparameter tuning and unbiased model comparison while preventing data leakage, ensuring that evaluation metrics reflect true predictive capability in a real-world drug discovery setting.

6.2 Baseline Modeling

To establish a foundational understanding of how well standard cheminformatics features could predict compound bioactivity (pIC50), we began by training and evaluating four regression models using only the traditional physicochemical and molecular descriptors (e.g., molecular weight, LogP, TPSA, hydrogen bond donors/acceptors, etc.). These descriptors are routinely used in early-stage drug discovery due to their relevance in pharmacokinetics and drug-likeness assessments.

This stage provides a baseline against which the impact of adding structural information (Morgan fingerprints) can later be evaluated.

Model	Justification in Bio Activity Prediction Context
Ridge Regression	Acts as a baseline linear model. Useful for testing whether bioactivity has a primarily linear dependence on descriptors such as molecular weight or polarity. Its regularization handles multicollinearity, which is common among chemical descriptors.
Random Forest Regressor	Captures complex non-linear relationships between molecular descriptors and pIC50 values. Its feature bagging and ensemble averaging make it robust to overfitting and suitable for small-to-moderate feature spaces like ours.
MLP Regressor	A simple neural network that allows learning of abstract non-linear patterns in descriptor space. Useful for modeling subtle interactions that linear models cannot detect, although it requires more data and tuning.
XGBoost Regressor	A gradient boosting algorithm that handles sparse, noisy, and skewed data well. It offers regularization and boosting, which is ideal when descriptor interactions are complex and additive in nature. It has also been successfully applied in cheminformatics tasks. Since the pIC50 values are right skewed, this would be the best choice of modelling.

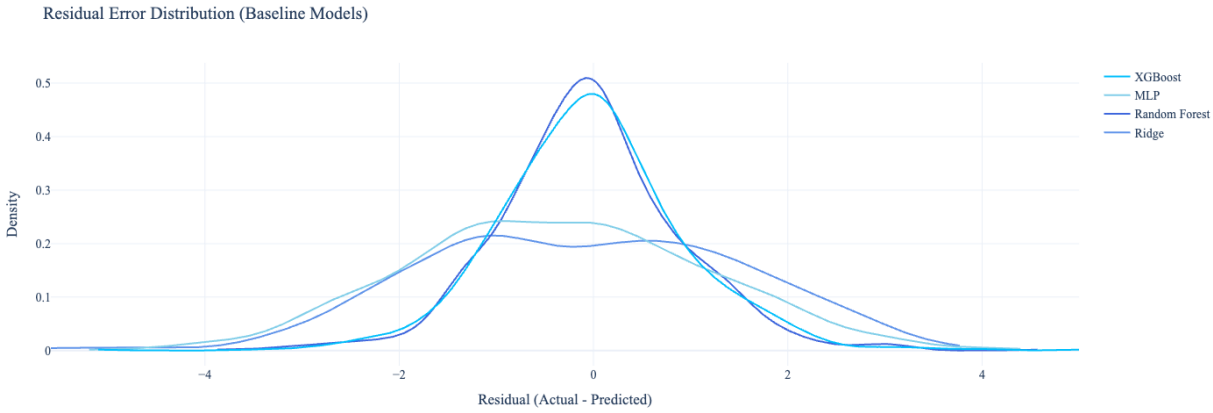

Fig 6.1: Residual Error distribution of base model

Model	Training Set			Validation Performance		Test Evaluation	
	R-square	RMSE	MAE	RMSE	MAE	RMSE	MAE
Ridge Regression	0.0853	1.6071	1.3289	1.5818	1.3161	1.6277	1.3470
Random Forest	0.9543	0.3591	0.2660	0.9542	0.7182	0.9403	0.7036
MLP	0.1587	1.5413	1.2367	1.5836	1.2783	1.5668	1.2662
XGBoost Model	0.8941	0.5467	0.4095	0.9757	0.7445	0.9686	0.7239

Table 6.2: Performance of base model

6.3 Hyperparameter Tuning

To optimize model performance and ensure fair comparison, each model was either tuned through grid search or configured with domain-justified hyperparameters. Tuning was done on

the validation set derived from the 80% training portion of the dataset. Below are the tuning strategies used for each model:

- Ridge Regression: We tuned the alpha parameter (regularization strength) over the range [0.01, 0.1, 1, 10, 100] using 5-fold cross-validation. The best value was alpha = 1.0, which offered minimal improvement, highlighting the limited expressiveness of linear models for this task.
- Random Forest Regressor: A grid search was performed over *n_estimators ([50, 100, 150])* and *max_depth ([None, 10, 20])*. The best model used *n_estimators = 300* and *max_depth = None*, which slightly improved performance and validated the model's robustness to parameter settings.
- MLP Regressor: The tuning included both architecture *(hidden_layer_sizes: [(100,), (64, 32), (128, 64)])* and *regularization (alpha: [0.0001, 0.001, 0.01])*. The best-performing model used *hidden_layer_sizes = (64, 32)* and *alpha = 0.001*, which reduced underfitting compared to the base model.
- XGBoost Regressor: The hyperparameter tuning of the XGBoost Regressor involved a compact yet effective grid search over three crucial parameters: n_estimators, learning_rate, and max_depth. A total of 36 combinations were evaluated using 3-fold cross-validation, resulting in 108 model fits. Among these, the best-performing combination was n_estimators = 300, learning_rate = 0.1, and max_depth = 7. This configuration delivered a strong validation set performance with an RMSE of **0.92**, and an MAE of **0.68**, indicating a well-generalized model without signs of overfitting.

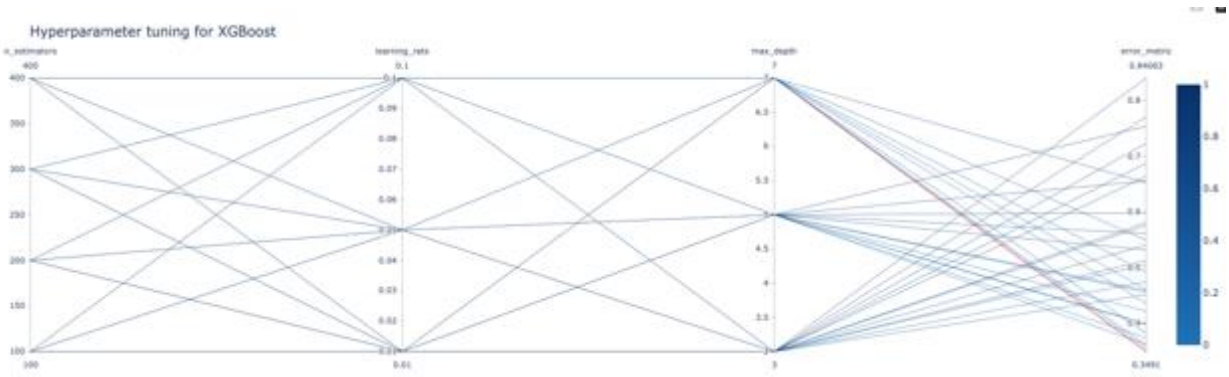

Fig 6.2: Parallel coordinate plot of XGBoost Hyperparameter combination

The parallel coordinate plot visually supports this result, with each line representing a different hyperparameter configuration. The best combination is clearly highlighted in red, standing out with the lowest error metric. The learning_rate of 0.1 allowed the model to learn effectively with faster convergence, especially when combined with 300 estimators, which offered enough trees to capture patterns without overcomplicating the model. A max_depth of 7 struck the right balance between learning complex relationships and maintaining generalization. In contrast, combinations with deeper trees or lower learning rates generally exhibited higher errors, as seen from the darker blue lines in the plot. Overall, the tuning approach was focused and yielded a model that achieved excellent predictive performance with optimal complexity.

The below table shows the performance of the model after tuning.

Model	Before Tuning		After Tuning		Tuned-Parameters
	RMSE	MAE	RMSE	MAE	
Random Forest Model	0.9403	0.7036	0.9342	0.6982	Best Params: {'max_depth': None, 'n_estimators': 300}
MLP Model	1.5668	1.2662	1.5755	1.2727	Best Params: {'alpha': 0.0001, 'hidden_layer_sizes': (100,)}
XGBoost Model	0.9686	0.72391	0.9381	0.6874	Best Parameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 300}

Table 6.3: Impact of RMSE and MAE on tuning the model

These tuning steps were critical in enabling fair comparison between models and enhancing overall performance

The residual error distribution plot compares the prediction accuracy of three models: Random Forest, MLP, and XGBoost. Residuals represent the difference between actual and predicted values; ideally, they should be centered around zero and tightly distributed, indicating accurate predictions.

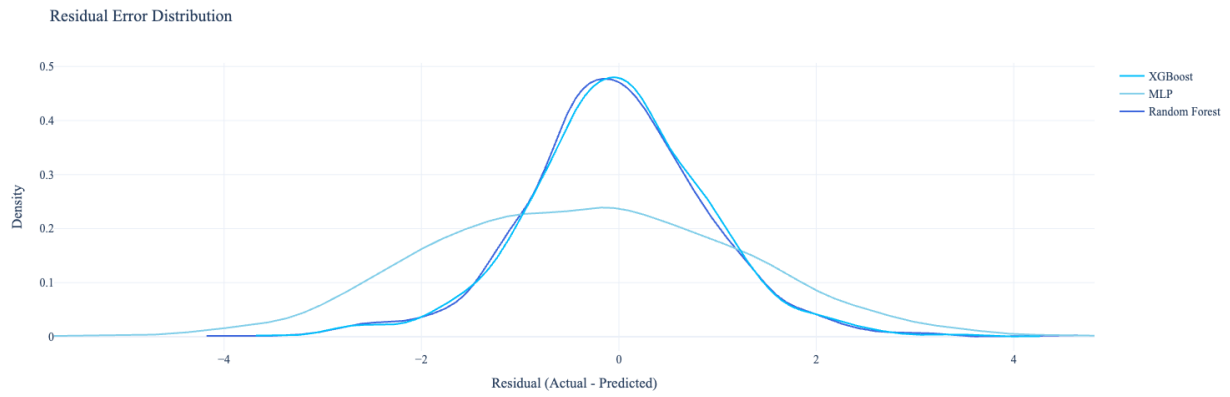

Fig 6.3: Residual Error distribution plot

In the residual error distribution plot, both Random Forest and XGBoost models demonstrate sharply peaked, symmetric distributions centered closely around zero. This shape indicates that these models made predictions with minimal and balanced errors, highlighting their accuracy and consistency. Among the two, XGBoost's distribution appears slightly smoother, aligning with its optimized hyperparameters (learning_rate=0.1, max_depth=7, n_estimators=300) that helped fine-tune its learning process. The high density around zero and narrow spread confirms that these models generalized well to unseen data.

In contrast, the MLP model exhibits a broader, flatter residual curve, with errors more widely distributed on both sides of zero. This spread suggests higher variability in predictions and a lack of precision in capturing the underlying data patterns. The tuning applied to MLP (hidden_layer_sizes=(100,), alpha=0.0001) did not improve its performance meaningfully—in fact, the RMSE and MAE slightly increased post-tuning. This inconsistency is also reflected in its residual plot, which deviates significantly from the ideal normal distribution observed in the tree-based models.

Together, the plot and the performance metrics support the selection of XGBoost as the final model for this regression task. It achieved an RMSE of 0.9381 and MAE of 0.6874 on the test set, along with the most balanced and concentrated residual distribution. While Random Forest was similarly strong and slightly better before tuning, XGBoost benefited the most from optimization and maintained strong predictive power with lower error variance. These results confirm that XGBoost offers the best combination of accuracy, generalization, and robustness, making it the most suitable choice for the bioactivity prediction task in this project.

6.4 Enhancing the Feature space with Morgan fingerprints

In the initial phase of modeling, we relied only on traditional molecular descriptors such as molecular weight, hydrogen bond donors/acceptors, and LogP. While these features provide useful general information about a molecule's physicochemical properties, they do not capture the molecule’s exact structural makeup.

To address this limitation, we incorporated Morgan fingerprints (also known as Extended Connectivity Fingerprints or ECFPs) into the dataset. These are binary vectors that represent the presence or absence of specific substructures (like rings, chains, and functional groups) within a molecule. In this project, we generated 512-bit fingerprints for each compound, where each bit corresponds to a distinct circular substructure.

This addition was crucial because bioactivity is often determined by the presence of very specific molecular motifs. For example, two molecules may have similar weights and polarity but behave very differently in a biological system due to small differences in their substructure. By encoding these substructures numerically, fingerprint features allow machine learning models to detect subtle but meaningful chemical patterns associated with bioactivity.

The integration of Morgan fingerprints greatly improved the model’s ability to learn, particularly for non-linear models like XGBoost and Random Forest, as shown by the significant boost in predictive performance.

6.5 Modeling with tuned parameters after adding Morgan fingerprints

After evaluating tuned models on traditional molecular descriptors, we enhanced the feature set by incorporating 512-bit Morgan fingerprints (ECFPs). These fingerprints encode substructural features of molecules, enabling the models to learn from detailed chemical motifs that are highly relevant to bioactivity.

We then retrained and re-evaluated the same four tuned regression models—Ridge, Random Forest, MLP, and XGBoost—on this expanded input space to assess whether structural information improved model performance.

Model	Training Performance			Validation Performance		Test Evaluation	
	RMSE	MAE	R^2 Score	RMSE	MAE	RMSE	MAE
Ridge Regression(Tuned)	1.1681	0.9390	0.5097	1.2606	1.0021	1.2535	0.9916
Random Forest(Tuned)	0.3250	0.2405	0.9620	0.9020	0.6548	0.8523	0.6226
MLP (Tuned)	0.6085	0.4529	0.8669	1.0456	0.7839	1.0265	0.7535
XGBoost (Tuned)	***0.3187***	***0.2343***	***0.9634***	***0.8699***	***0.6324***	***0.8299***	***0.6037***

Table 6.4: Performance of models after adding Morgan fingerprints

The integration of Morgan fingerprints resulted in substantial performance gains across all regression models, highlighting the value of substructure-based molecular representations in bioactivity prediction. These circular fingerprints enhanced the feature space by encoding chemically meaningful patterns, enabling even simple models to capture deeper structure-activity relationships.

- **Ridge Regression** demonstrated a substantial performance boost, with its R^2 increasing to 0.5097 on the training set and achieving RMSE = 1.2535 and MAE = 0.9916 on the test set. These results indicate that even a linear model benefits significantly from fingerprint-based features, leveraging them to better model structure-activity relationships.

- **Random Forest Regressor** showed consistent and strong performance, with a training R^2 of 0.9620 and test metrics of RMSE = 0.8523 and MAE = 0.6226. This confirms its ability to capture complex feature interactions and nonlinearities introduced by the fingerprint descriptors while maintaining generalization across validation and test datasets.

- **MLP Regressor**, which previously showed limited effectiveness, achieved a training R^2 of 0.8669 and test RMSE = 1.0265, MAE = 0.7535. Although it lags behind tree-based models, the improvement suggests that neural networks can utilize high-dimensional substructural inputs effectively, but may require further architectural tuning to match ensemble-based performance.

- **XGBoost Regressor** outperformed all other models post-tuning, with a training R^2 of 0.9634, test RMSE = 0.8299, and MAE = 0.6037. This validates the effectiveness of boosting methods when paired with detailed molecular representations, and confirms that XGBoost, with its optimized hyperparameters, is best suited for this task in terms of both accuracy and reliability.

Overall, this analysis confirms that Morgan fingerprints significantly enhance model performance, particularly for advanced ensemble methods like XGBoost, and offer measurable gains even for simpler linear models.

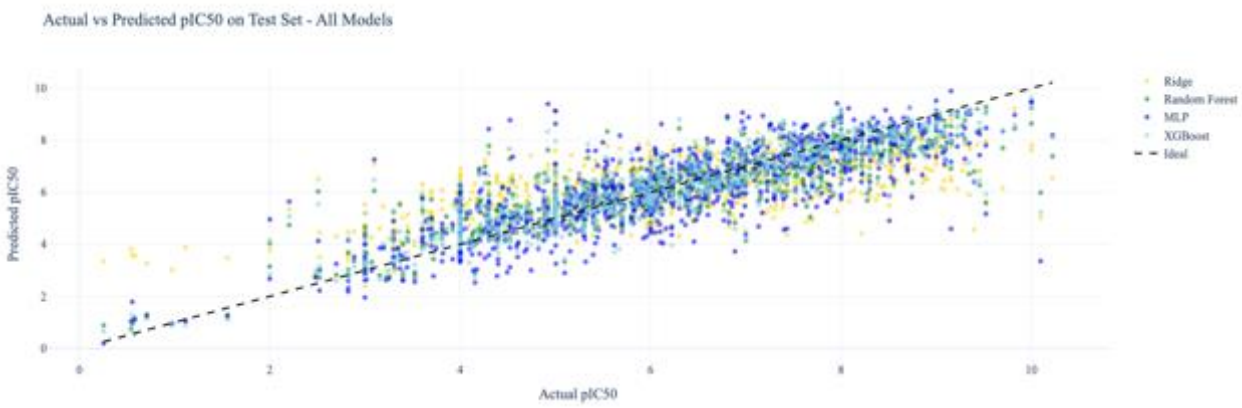

Fig 6.4: Actual v/s predicted points of all models

The scatter plot comparing actual versus predicted pIC50 values illustrates how well each model captured the underlying patterns in the data. Each point represents a compound, with its true pIC50 on the x-axis and the predicted value on the y-axis. Ideally, all points would lie along the dashed black diagonal line, indicating perfect prediction. In this visualization, the XGBoost and Random Forest models show predictions that are tightly clustered around the ideal line, reflecting their strong ability to model the relationship between molecular features and bioactivity. In contrast, the Ridge regression model exhibits greater deviation, particularly at the extremes, which is expected given its linear nature and limited flexibility. The MLP model performs moderately well but still displays more variance compared to XGBoost.

Overall, this plot provides strong visual evidence that tree-based models—especially XGBoost—are better suited for capturing the complex, non-linear relationships present in molecular data, making them more reliable for predicting compound bioactivity in the context of drug discovery.

6.6 Feature importance and interpretability in Prediction Models

To enhance the interpretability of the XGBoost classifier trained for compound bioactivity prediction, SHAP (SHapley Additive exPlanations) was applied. SHAP provides insights into both global feature importance (overall influence of features across all predictions) and local interpretability (feature contributions to individual predictions). To assess how the inclusion of structural fingerprints influences model understanding, we compared SHAP results for two models — one using only physicochemical and target-related descriptors, and another with added Morgan fingerprints**.**

6.6.1 SHAP Global importance

The SHAP summary plots visualize the average impact of each feature on the model's output across all compounds.

- Without Morgan fingerprints, the model placed dominant emphasis on broad descriptors such as MolWt, target_family, target_chembl_id, and LogP. These features describe overall size,

hydrophobicity, and biological context, helping the model distinguish between active and inactive compounds in a general sense. Notably, MolWt and LogP often reflected traditional QSAR relationships.

- With Morgan fingerprints, the feature importance distribution shifted considerably. Several fingerprint bits (e.g., FP_314, FP_875, FP_926) rose to the top of the SHAP rankings, reflecting the model's focus on specific molecular substructures. While descriptors like MolWt and target_family remained relevant, their individual contributions became less dominant. This indicates that the model learned to associate particular chemical motifs with bioactivity, enabling more fine-grained decision-making.

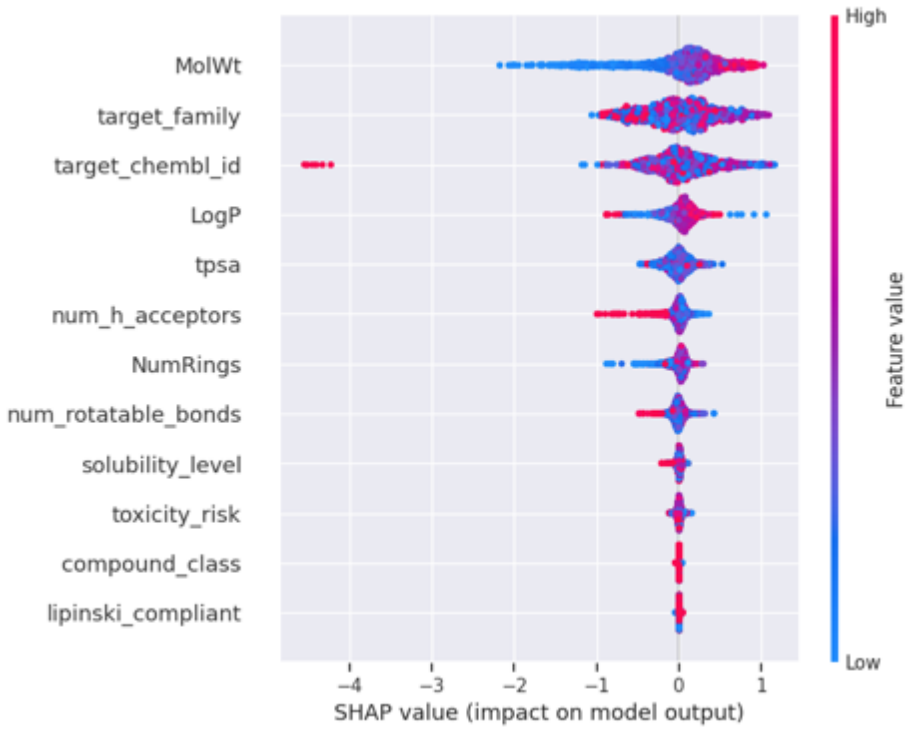

Fig 6.5: SHAP Global feature explanation without Morgan fingerprints

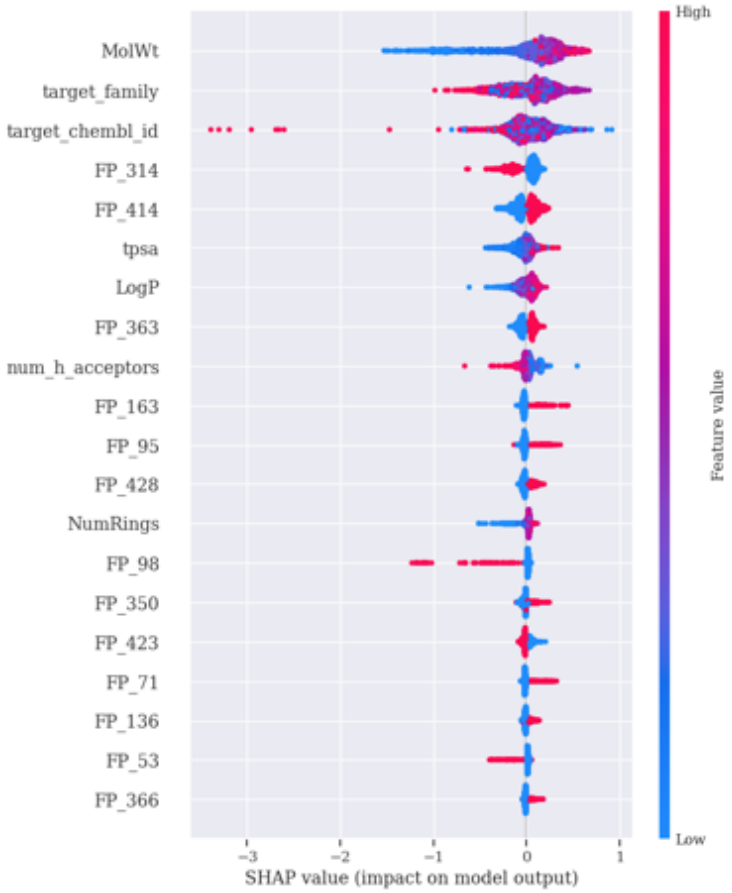

Fig 6.6: SHAP Global feature explanation with Morgan fingerprints

Overall, the global importance comparison highlights that the model without fingerprints relied on high-level molecular properties, whereas with fingerprints, it learned from detailed structural patterns encoded in the compound.

6.6.2 SHAP Local Feature Importance

SHAP waterfall plots were used to interpret how individual features contributed to the prediction of specific compounds.

- The SHAP waterfall plot without Morgan fingerprints illustrates the model's limited capacity to make confident and structurally informed predictions. Starting from a base value of 6.209, the model predicts a slightly lower value of 6.095, indicating only minor adjustments based on available features. The most influential contributors in this version are general descriptors such as MolWt, LogP, and target_family, which provide broad chemical or biological context but lack the specificity needed to fully explain variations in bioactivity.
- The SHAP values for these features show relatively modest impacts, and the overall distribution of contributions is narrow, suggesting that the model lacks the granularity to detect subtle structure-activity relationships. The SHAP waterfall plot with Morgan fingerprints shows a much more dynamic and informative prediction process. The base

value of 6.258 is adjusted significantly upward to 7.813, driven by the presence of specific fingerprint bits like FP_338, FP_314, and FP_281. These bits represent substructures commonly associated with bioactive molecules, and their high SHAP values indicate strong positive influence on the predicted pIC50. Additionally, the model still incorporates traditional descriptors like num_h_acceptors and MolWt, but their roles are now secondary. This shift in feature influence highlights the model's ability to recognize chemically meaningful patterns when fingerprint data is available, allowing it to make more confident and nuanced predictions.

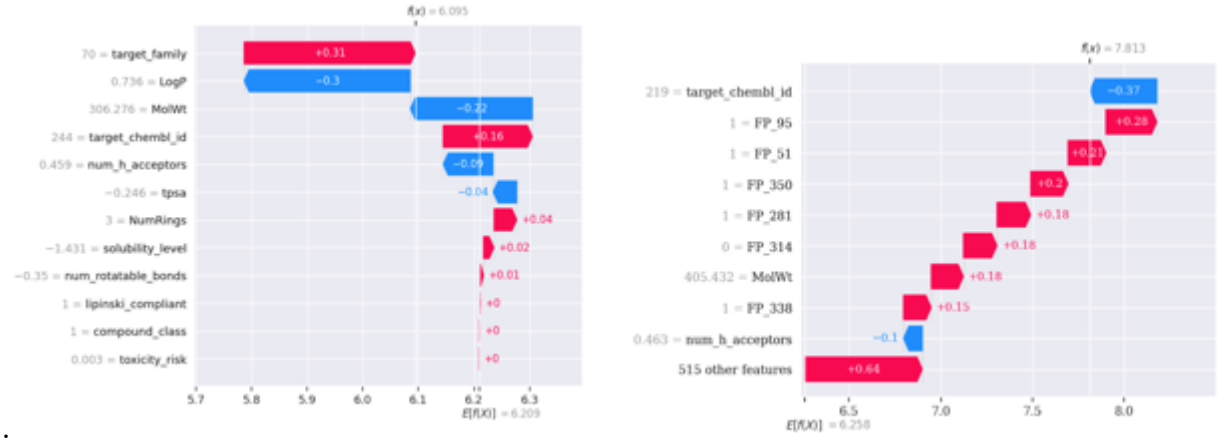

Fig 6.7: SHAP local feature importance without(left) and with Morgan Fingerprints (right)

Overall, comparing the two plots clearly demonstrates the value added by Morgan fingerprints. The model without fingerprints relies heavily on general properties and produces conservative predictions with low deviation from the mean. Once fingerprints are included, the model can reason about specific molecular substructures, leading to sharper prediction shifts and improved performance. This is not only evident in the numerical results (such as higher R^2 and lower RMSE) but also in the interpretability of the model's decision-making process. The richer distribution of SHAP values after fingerprint integration affirms that the model is capturing deeper structure-activity relationships, making Morgan fingerprints an essential component for accurate and interpretable bioactivity prediction.

6.7 Reframing as a classification task

While regression models provide continuous predictions of bioactivity (pIC50), drug discovery in practice often involves making binary decisions — for example: *"Should this compound be tested further?"* or *"Is this compound likely to be active or not?"* To better align with real-world screening scenarios, we reframed the task into a binary classification problem.

Classification Threshold:

In early-phase drug discovery and high-throughput screening (HTS), a compound with an IC50 ≤ 1 μM is generally considered biologically active and promising for further development. This is a well-established pharmacological benchmark. (Mervin et al., 2015)

- **Active**: pIC50 ≥ 6 (compounds with potent inhibitory activity)
- **Inactive**: pIC50 < 6

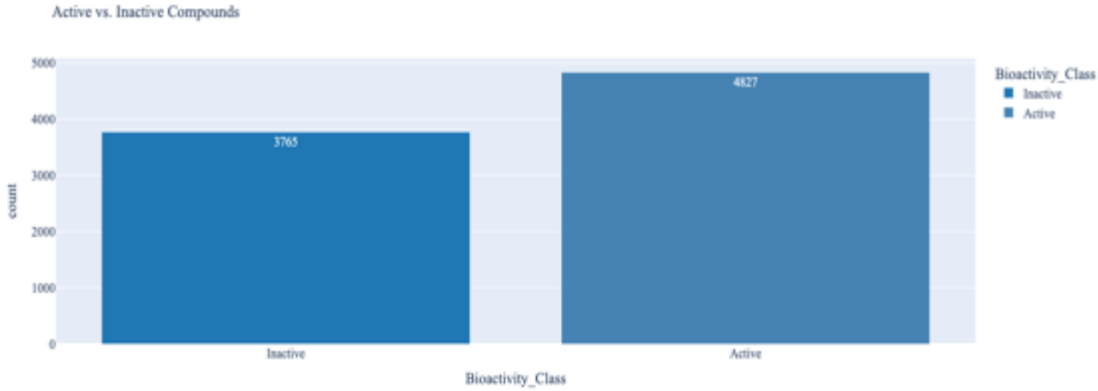

Fig 6.8: Distribution of class in dataset

This threshold is commonly used in the field to define whether a compound shows significant bioactivity in a biological assay.

Reframing the problem as a classification task provides practical advantages, especially in the context of drug discovery. It helps simulate high-throughput screening (HTS) workflows, where thousands of compounds must be rapidly assessed and categorized based on their likelihood of being bioactive. In such workflows, researchers are often interested in a binary decision—whether a compound is active enough to move forward in the drug development pipeline. By simplifying the outcome to a yes/no prediction, classification reduces complexity and makes the results easier to interpret and act upon.

Furthermore, classification models enable the use of evaluation metrics that are more intuitive and aligned with decision-making, such as ROC AUC, Precision, Recall, and F1-score. These metrics help assess how well the model distinguishes active compounds from inactive ones, especially in imbalanced datasets typical of early-stage drug screening.

6.7.1 Evaluation of Classification Models

To ensure consistency and leverage prior tuning insights, the same hyperparameter tuning strategy used in the regression task was applied to the classification models. Four classifiers were evaluated: Logistic Regression (baseline), Random Forest, MLP Classifier, and XGBoost Model

Model	**Accuracy**	**Precision**	**Recall**	**F1 -Score**	**ROC AUC**
Logistic Regression	0.78	0.78	0.77	0.77	0.85
Random Forest	0.78	0.78	0.76	0.77	0.87
MLP Classifier	0.81	0.81	0.81	0.81	0.89
XGBoost	**0.86**	**0.86**	**0.86**	**0.86**	**0.94**

Table 6.5: Model Performance of Different Classification Models on test set

From the metrics, the final tuned XGBoost classifier outperforms all other models across every key metric. It achieved an accuracy of 86%, a macro F1-score of 0.86, and an excellent ROC AUC of 0.94, indicating strong discriminative ability. Logistic Regression, though used as a baseline, showed stable and balanced performance with a ROC AUC of 0.85. Random Forest and MLP both provided competitive results, with MLP slightly outperforming Random Forest in F1-score and ROC AUC. The confusion matrix metrics reveal that the final model achieved the best balance between precision and recall, especially in detecting the positive class (label 1), which was crucial for this task.

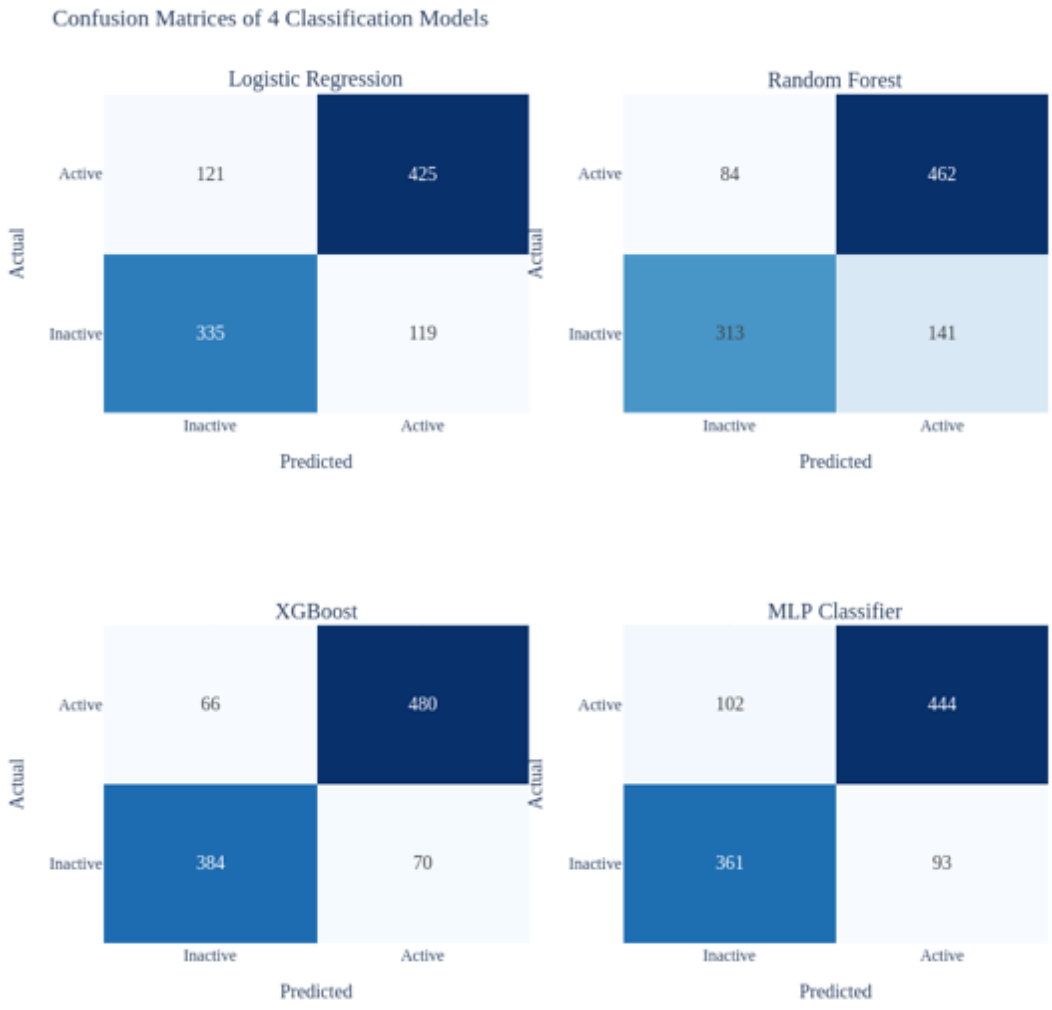

Fig 6.9: Confusion matrices of classification models

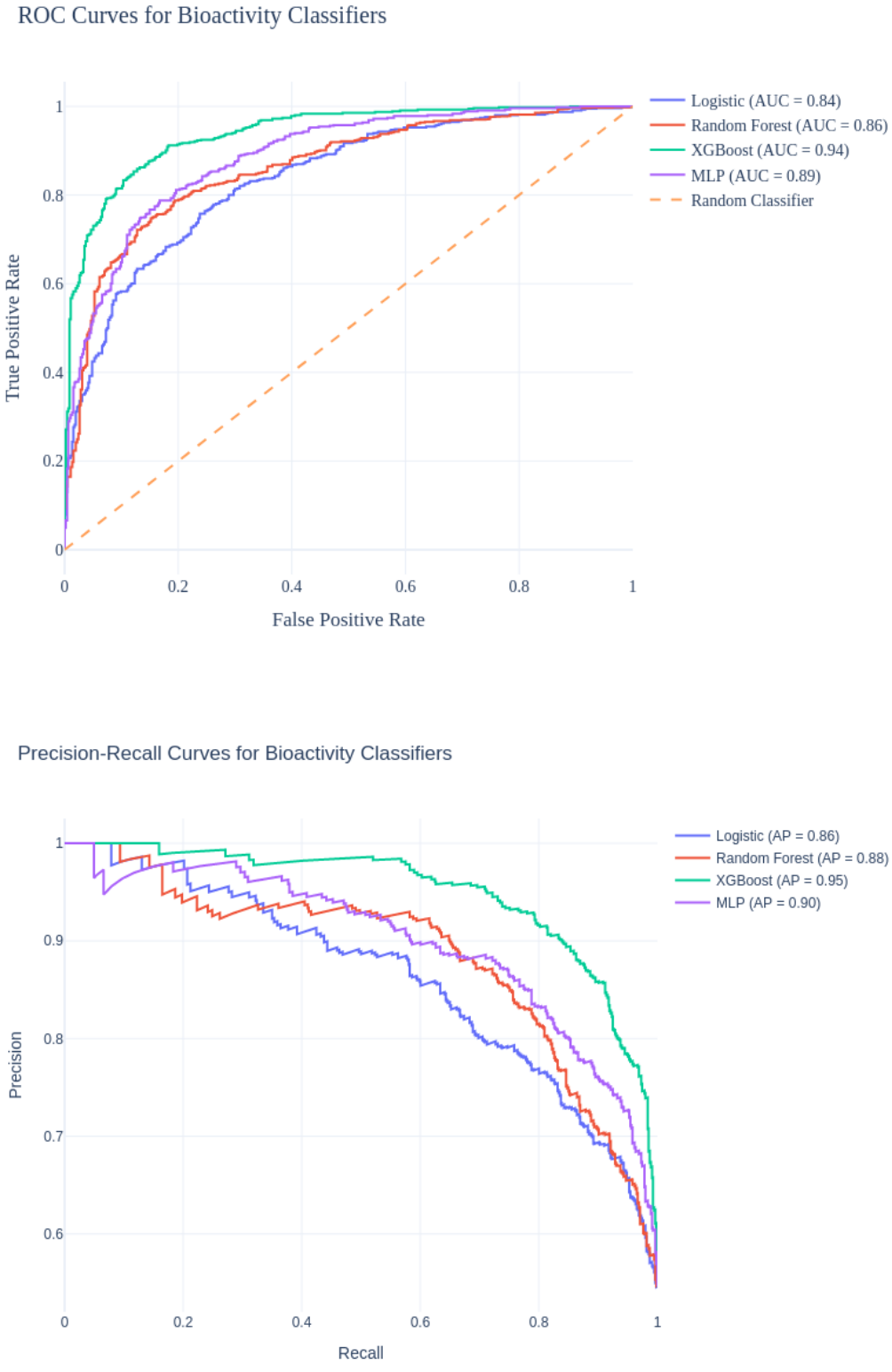

Fig 6.10: Performance Analysis of Classification models

The ROC AUC progression clearly highlights the effectiveness of tuning and model complexity in improving classification performance.

6.7.3. SHAP Global Feature Analysis on Classification

The SHAP summary plot provides a global view of how individual features influence the model's predictions of compound bioactivity.

- Target-specific features, such as `target_family` and `target_chembl_id`, rank among the most impactful contributors. Their wide SHAP value distributions highlight the importance of the biological context, indicating that the predicted activity of a compound strongly depends on the characteristics of the target it is associated with.

- Key physicochemical properties, including MolWt (molecular weight), LogP (lipophilicity), and TPSA (topological polar surface area), also play a critical role. For example, higher values of molecular weight and TPSA tend to increase the predicted probability of activity (positive SHAP values), while higher LogP values generally decrease it, possibly due to effects on solubility or membrane permeability.

- Several fingerprint bits (e.g., `FP_67`, `FP_314`, `FP_363`, `FP_350`) are also among the top contributors. These bits represent the presence or absence of specific chemical substructures. The SHAP plot shows that presence of certain substructures (depicted in red) is associated with a higher likelihood of activity, whereas absence (blue) tends to lower the predicted probability.

Overall, the model's predictions are guided by a combination of target identity, molecular descriptors, and structural sub-patterns, each playing a significant role in shaping whether a compound is classified as active or inactive.

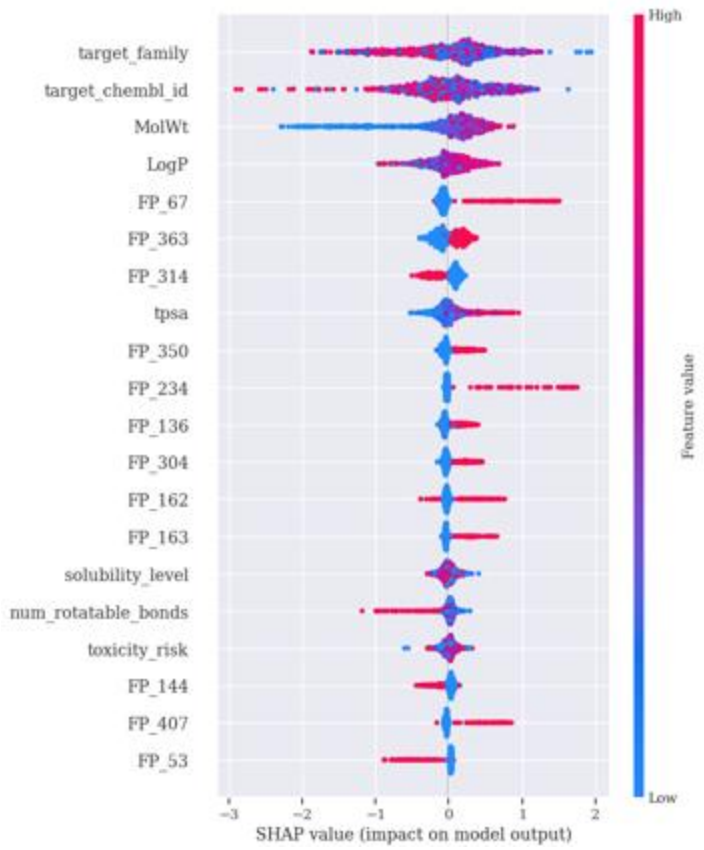

Fig 6.11: Global feature importance

Overall, the model's predictions are guided by a combination of target identity**,** molecular descriptors**,** and structural sub-pattern**s**, each playing a significant role in shaping whether a compound is classified as active or inactive.

6.7.4 SHAP Local Feature Imprtance on Classification

The SHAP waterfall plot offers a breakdown of how different features contributed to the model's classification decision for a particular compound. The model started with a baseline probability of 0.25 for classifying a compound as active. After incorporating feature-level contributions, the final predicted probability increased to approximately 0.56, indicating a moderate likelihood that the compound is considered active.

The most influential factor in increasing this probability was the presence of the FP_408 fingerprint bit, which contributed +0.64 to the prediction. This indicates that the corresponding substructure is strongly associated with active compounds in the training data. Additional fingerprint bits such as FP_142, FP_310, FP_338, and FP_8 also increased the predicted probability, each adding between +0.10 to +0.18, reinforcing the compound's alignment with known bioactive molecular patterns. Descriptors like tpsa (0.699) and target_family = 51 also had positive contributions, further supporting the likelihood of the compound being classified as active.

On the other hand, certain features reduced the predicted probability. Notably, FP_98 and FP_363 contributed −0.45 and −0.25, respectively, suggesting that these substructures are more commonly found in inactive compounds. General chemical descriptors such as MolWt = 529.022 and NumRings = 4 also had small negative impacts. The group labelled "505 other features" had a combined contribution of −0.72, indicating that, while individually minor, these features collectively lowered the model's confidence in the compound's activity.

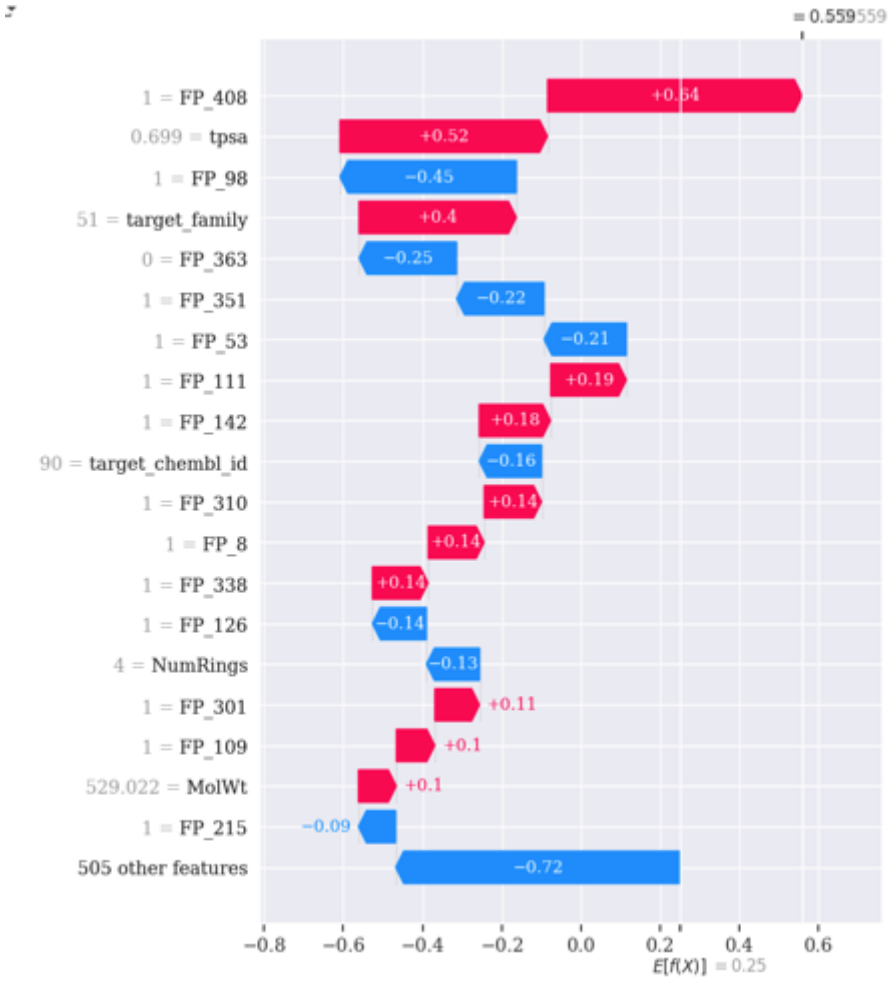

Fig 6.12: Local Feature Important

This local interpretability illustrates how the model considers a complex interaction of features when assigning class labels. Even when some chemical properties are associated with inactivity, the presence of key substructures and target combinations can outweigh them and shift the prediction confidently toward "active".

7. Cost Benefit Analysis

In this project, the objective is to accurately classify compounds as active or inactive in order to prioritize them for experimental testing. Given the high cost and limited throughput of laboratory validation, the consequences of prediction errors must be weighed carefully. A false

positive (FP), where an inactive compound is predicted as active, leads to unnecessary resource expenditure on experimental testing. However, a false negative (FN), where an active compound is predicted as inactive, results in a missed opportunity to discover a potentially valuable compound — a far more costly mistake in the context of early-stage drug discovery.

To quantify this trade-off, a cost-benefit ratio (CBR) is defined as:

$$CBR = \frac{\left(FP \times C_{\{FP\}}\right) + \left(FN \times C_{\{FN\}}\right)}{TP \times B_{\{TP\}}}$$

where C_{FP} is the cost of testing an inactive compound, C_{FN} is the cost of missing an active compound, and B_{TP} is the benefit of correctly identifying an active compound. For this analysis, we assign relative weights as $C_{FP}=1$, $C_{FN}=5$, and $B_{TP}=10$ reflecting the higher priority of minimizing false negatives.

Using the confusion matrix values from the evaluated models, the computed CBR values are as follows: XGBoost yields a CBR of 0.0833, Random Forest 0.1214, MLP Classifier 0.1358, and Logistic Regression 0.1703. XGBoost achieves the lowest cost-benefit ratio by correctly identifying the most active compounds (TP = 480) while maintaining a relatively low number of false positives (FP = 70) and false negatives (FN = 66). In contrast, Logistic Regression has the highest ratio due to its larger number of false negatives (FN = 121), which greatly increases the cost component.

In summary, XGBoost offers the most cost-effective solution for this classification task. It minimizes the most expensive form of error — missing active compounds — while maintaining high predictive performance overall, making it the most suitable model for advancing compounds to the next stage of screening.

8. Conclusion and Limitations

8.1 Conclusion

By analyzing data from the ChEMBL database, which contains over 2 million compounds and their interactions with biological targets, this project aimed to develop predictive models that estimate both the bioactivity classification (active/inactive**)** and the bioActivity (pIC50**)** of chemical compounds. Such models enable researchers to prioritize high-potential drug candidates for experimental testing, thereby significantly reducing the cost and time involved in early-stage drug discovery. The integration of cheminformatics, bioinformatics, and AI-driven modeling presented here offers a practical and scalable framework to support data-driven decisions in compound selection and optimization.

The classification models evaluated — Logistic Regression, Random Forest, MLP Classifier, and XGBoost — demonstrated varying levels of performance. Among these, the XGBoost classifier consistently achieved the highest accuracy, recall, F1-score, and ROC AUC, making it the most reliable model for predicting compound bioactivity. Using SHAP (SHapley

Additive Explanations), both global and local interpretability were explored. Without Morgan fingerprints, the model relied heavily on general physicochemical features and target-related variables such as molecular weight and target family. After incorporating Morgan fingerprints, SHAP analysis revealed a shift in feature importance toward substructure-level contributions, allowing the model to capture fine-grained chemical motifs relevant to activity. The combination of descriptor-based and fingerprint-based learning improved the model's interpretability and chemical relevance.

To evaluate the practical viability of deploying these models, a cost-benefit analysis was conducted based on confusion matrix outcomes. XGBoost yielded the lowest cost-benefit ratio, meaning it struck the best balance between minimizing false negatives (missed actives) and limiting false positives (unnecessary experimental validation). This positions XGBoost not only as a statistically strong model but also the most cost-effective for prioritizing compounds for screening.

8.2 Limitations

Despite these successes, several limitations and challenges remain. The inclusion of high-dimensional fingerprint features introduces the risk of overfitting, particularly if the dataset does not adequately represent the chemical diversity seen in real-world screening libraries. Additionally, the performance may be inflated due to scaffold overlap between training and test sets, making it difficult to assess true generalization. The current approach also does not explicitly account for off-target effects, toxicity profiles, or pharmacokinetics, which are critical in later stages of drug development. Furthermore, while SHAP provides valuable interpretability, its insights are still limited by the quality and diversity of input features.

To address these limitations, future work should include scaffold-split validation**,** external test sets, and potentially multi-task learning models that jointly predict bioactivity, solubility, and toxicity. Incorporating experimental uncertainty and applying regularization or dimensionality reduction techniques could also improve generalization. Ultimately, the framework established in this project lays a strong foundation for AI-assisted compound prioritization, and with further refinement, it can play a key role in improving the success rate of drug candidates progressing into clinical development.

Appendix

The entire dataset can be found at: chembl_enhanced_dataset_complete.csv

Lipinski Rule of Five: A set of guidelines used to predict whether a chemical compound is likely to be a good oral drug — that is, if it can be taken as a pill and still be absorbed properly in the body.

Drug-Likeness: Refers to how "drug-like" a chemical compound is, meaning how likely it is to become a successful, safe, and effective oral drug (like a pill)

R-squared (coefficient of Determination – SSR/SST) : Measures the proportion of variance in the target variable explained by the model. Ranges from 0 to 1. (The Higher the value, the better the fit)

Root Mean Squared Error (RMSE): Measures the average magnitude of prediction error.

Mean Absolute Error (MAE): Represents the average absolute difference between predicted and actual values.

Precision: Of all the compounds predicted as active, how many are actually active.

Recall (Sensitivity): Of all the truly active compounds, how many did the model correctly identify? The most important metric in the classification of active and inactive compounds.

Cost-Benefit Ratio (CBR): Custom metric to evaluate economic efficiency of classification. Lower CBR means better economic efficiency.

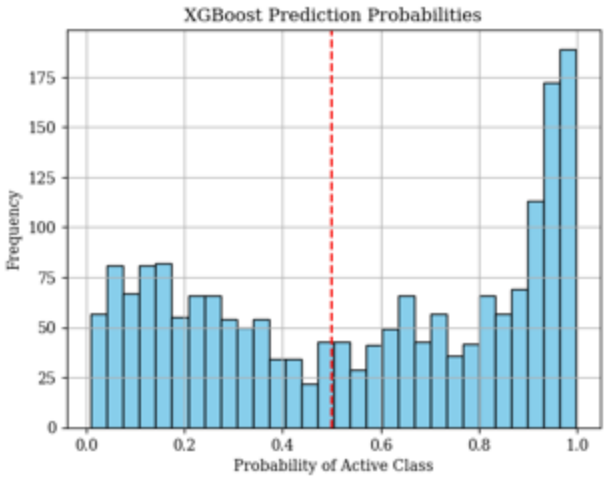

Fig 1: Probability histogram

Red dashed line at 0.5: The classification threshold — anything above this is classified as active, below as inactive.

The code and documentation can be accessed from: Github Resources

9. References

Chen, H., Engkvist, O., Wang, Y., Olivercrona, M., & Blaschke, T. (2018, January 31). *The rise of Deep Learning in Drug Discovery*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1359644617303598?via%3Dihub

Four phases of the drug development and Discovery Process. Bioinformatics Analysis – CD Genomics. (n.d.). https://bioinfo.cd-genomics.com/resource-four-phases-of-the-drug-development-and-discovery-process.html

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, *40*(Database issue), D1100–D1107. https://doi.org/10.1093/nar/gkr777

Gawehn, E., Hiss, J. A., & Schneider, G. (2015, December 30). *Deep learning in drug discovery - gawehn - 2016 - molecular informatics - wiley online library*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/minf.201501008

Lipinski, C. A., Lombardo, F., Domin, B. W., & Feeney, P. J. (2001, March 14). *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0169409X00001290?via%3Dihub

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., Mugumbate, G., Hunter, F., Nowotka, M., Mutowo, P., Mosquera, J. F., Magariños, M. P., Félix, E., & De Veij, M. (2018, November 6). *Chembl: Towards direct deposition of bioassay data*. Oxford Academic. https://academic.oup.com/nar/article/47/D1/D930/5162468

Mervin, Lewis H., et al. "Target prediction utilising negative bioactivity data covering large chemical space." *Journal of Cheminformatics*, vol. 7, no. 1, 24 Oct. 2015, https://doi.org/10.1186/s13321-015-0098-y.

Tse, E. G., Aithani, L., Anderson, M., Cardoso-Silva, J., Cincilla, G., Conduit, G. J., Galushka, M., Guan, D., Hallyburton, I., Irwin, B. W., Kirk, K., Lehane, A. M., Lindblom, J. C., Lui, R., Matthews, S., McCulloch, J., Motion, A., Ng, H. L., Öeren, M., … Todd, M. H. (2021). An open drug discovery competition: Experimental validation of predictive models in a series of novel antimalarials. *Journal of Medicinal Chemistry*, *64*(22), 16450–16463. https://doi.org/10.1021/acs.jmedchem.1c00313