DRUG DISCOVERY AND MACHINE LEARNING

Bioinformatics and Biostatistics Report

Contents

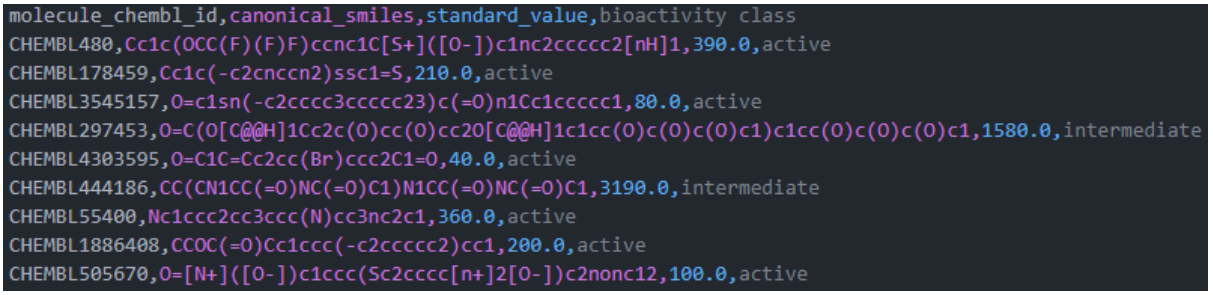
1. Introduction

In this study, we aim to analyze the bioactivity of compounds using IC50 values derived from the ChEMBL database. We classify these compounds into bioactivity categories (active, inactive, intermediate) and calculate molecular descriptors like Molecular Weight (MW), LogP, Hydrogen Bond Donors, and Hydrogen Bond Acceptors. Additionally, we perform statistical analysis and construct a machine learning model (QSAR) to predict bioactivity using molecular fingerprints.

2. Data Preparation

The data was collected from ChEMBL for the target receptor Coronavirus and the retrieve only bioactivity data for Replicase polyprotein 1ab (CHEMBL4523582). Bioactivity data was filtered based on IC50 values, seeing that these values indicate the potency of the drug, the lower the value the better the potency. Ideally the standard value should be as low as possible so that the inhibitory concentration at 50% would have a low concentration, meaning that to elicit(cause) 50% of the inhibition of a target protein there would be a need for a lower concentration of the drug. Compounds having values of less than 1000 nM will be considered to be active while those greater than 10,000 nM will be considered to be inactive. As for those values between 1,000 and 10,000 nM will be referred to as intermediate.

The following image shows a preview of the bioactivity_curated_data.csv that resulted from the data preparation process containing the compound's molecule chembl id, canonical smiles, IC50 values and their classification as either active, inactive or intermediate.



```
molecule_chembl_id,canonical_smiles,standard_value,bioactivity class
CHEMBL480,Cc1c(OCC(F)(F)F)ccnc1C[S+]([O-])c1nc2ccccc2[nH]1,390.0,active
CHEMBL178459,Cc1c(-c2cnccn2)ssc1=S,210.0,active
CHEMBL3545157,O=c1sn(-c2cccc3ccccc23)c(=O)n1Cc1ccccc1,80.0,active
CHEMBL297453,O=C(O[C@@H]1Cc2c(O)cc(O)cc2O[C@@H]1c1cc(O)c(O)c(O)c1)c1cc(O)c(O)c(O)c1,1580.0,intermediate
CHEMBL4303595,O=C1C=Cc2cc(Br)ccc2C1=O,40.0,active
CHEMBL444186,CC(CN1CC(=O)NC(=O)C1)N1CC(=O)NC(=O)C1,3190.0,intermediate
CHEMBL55400,Nc1ccc2cc3ccc(N)cc3nc2c1,360.0,active
CHEMBL1886408,CCOC(=O)Cc1ccc(-c2ccccc2)cc1,200.0,active
CHEMBL505670,O=[N+]([O-])c1ccc(Sc2cccc[n+]2[O-])c2nonc12,100.0,active
```

Picture 1: bioactivity_curated_data.csv preview.

3. Molecular Descriptors

Following the data preparation process we now have a dataset that consists of molecule names and their corresponding smiles notation(chemical structure). Using this information we can calculate the Lipinski descriptors.
Christopher Lipinski, a scientist at Pfizer, came up with a set of rule-of-thumb for evaluating the drug likeness of compounds. Such drug likeness is based on the Absorption, Distribution, Metabolism and Excretion (ADME) that is also known as the pharmacokinetic profile. Lipinski analyzed all orally active FDA-approved drugs in the formulation of what is to be known as the Rule-of-Five or Lipinski's Rule.

The Lipinski's Rule stated the following:

Molecular weight < 500 Dalton
Octanol-water partition coefficient (LogP) < 5
Hydrogen bond donors < 5
Hydrogen bond acceptors < 10

The Lipinski calculation takes in the smiles notation. That way the chemical structure is used to give the function the chemical information necessary to complete its computation.To allow IC50 data to be more uniformly distributed,

we will convert IC50 to the negative logarithmic scale which is essentially -log10(IC50).

The following image shows a preview of the results of these calculations, displaying a table containing the compound's molecule chembl id, bioactivity class and the four Lipinski descriptors as well as the pIC50 values.

	molecule_chembl_id	bioactivity class	MW	LogP	NumHDonors	NumHAcceptors	pIC50
0	CHEMBL480	active	369.368	3.51522	1.0	4.0	6.408935
1	CHEMBL178459	active	226.351	3.30451	0.0	5.0	6.677781
2	CHEMBL3545157	active	334.400	3.26220	0.0	5.0	7.096910
4	CHEMBL4303595	active	237.052	2.22770	0.0	2.0	7.397940
6	CHEMBL55400	active	209.252	2.55240	2.0	3.0	6.443697
...	...	...	...	...	...	...	...
1128	CHEMBL5279748	active	485.507	0.99528	3.0	5.0	7.433916
1129	CHEMBL5283975	active	485.507	0.99368	3.0	5.0	7.262489
1130	CHEMBL5266964	active	507.657	0.50048	3.0	6.0	7.641494
1131	CHEMBL5286307	active	499.534	1.38378	3.0	5.0	7.743282
1132	CHEMBL5282079	active	535.589	0.85788	3.0	6.0	7.649364

Picture 2: Table containing the compound's molecule chembl id, bioactivity class and the four Lipinski descriptors as well as the pIC50 values.

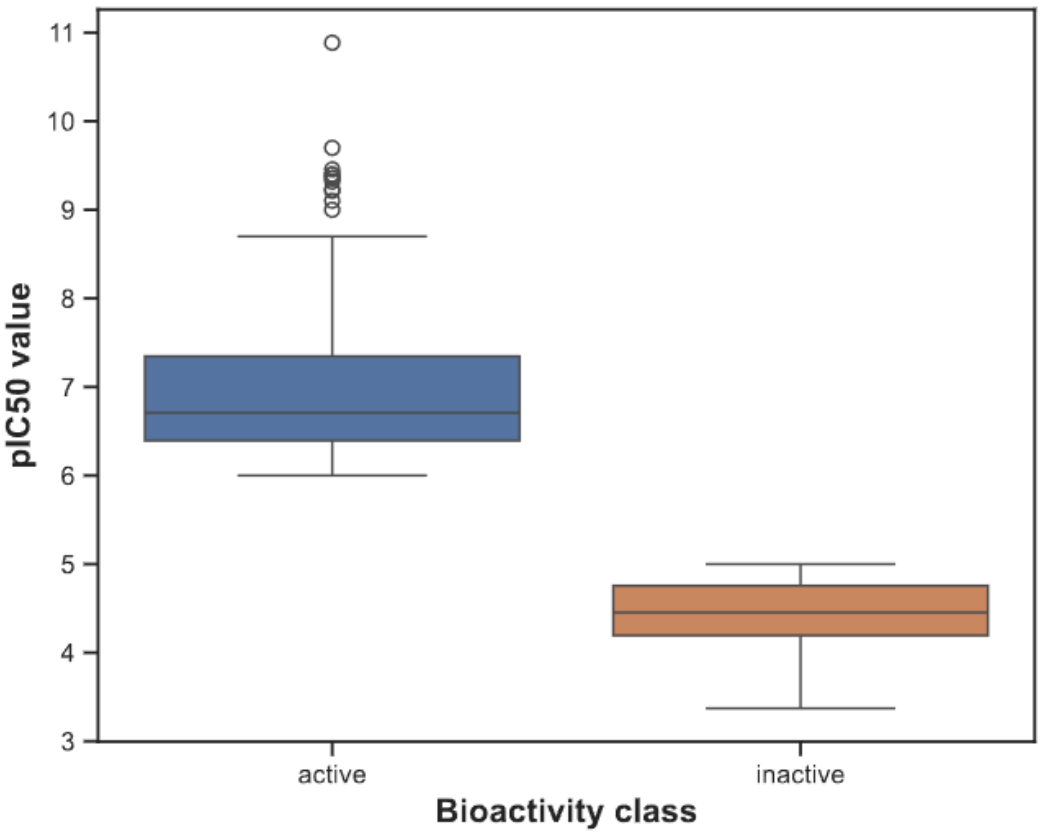
4. Statistical analysis

In this step we utilized The Mann-Whitney U Test (also known as the Wilcoxon rank-sum test), a non-parametric test used to determine whether there is a statistically significant difference between two independent groups, to compare the Lipinski descriptors and pIC50 values between bioactivity classes.

Performing the Mann-Whitney U Test for each descriptor helps identify molecular properties that significantly differ between two groups (e.g., active vs. inactive compounds). This is useful in bioactivity analysis, cheminformatics, and drug discovery to understand which molecular features contribute to the activity of compounds.

For example, if the Mann-Whitney U Test shows that **MW** has a significant difference between active and inactive compounds, it might indicate that molecular weight plays a role in bioactivity. Similar interpretations can be made for **LogP** (hydrophobicity), **H-bond acceptors (HBA)**, or **H-bond donors (HBD)**.

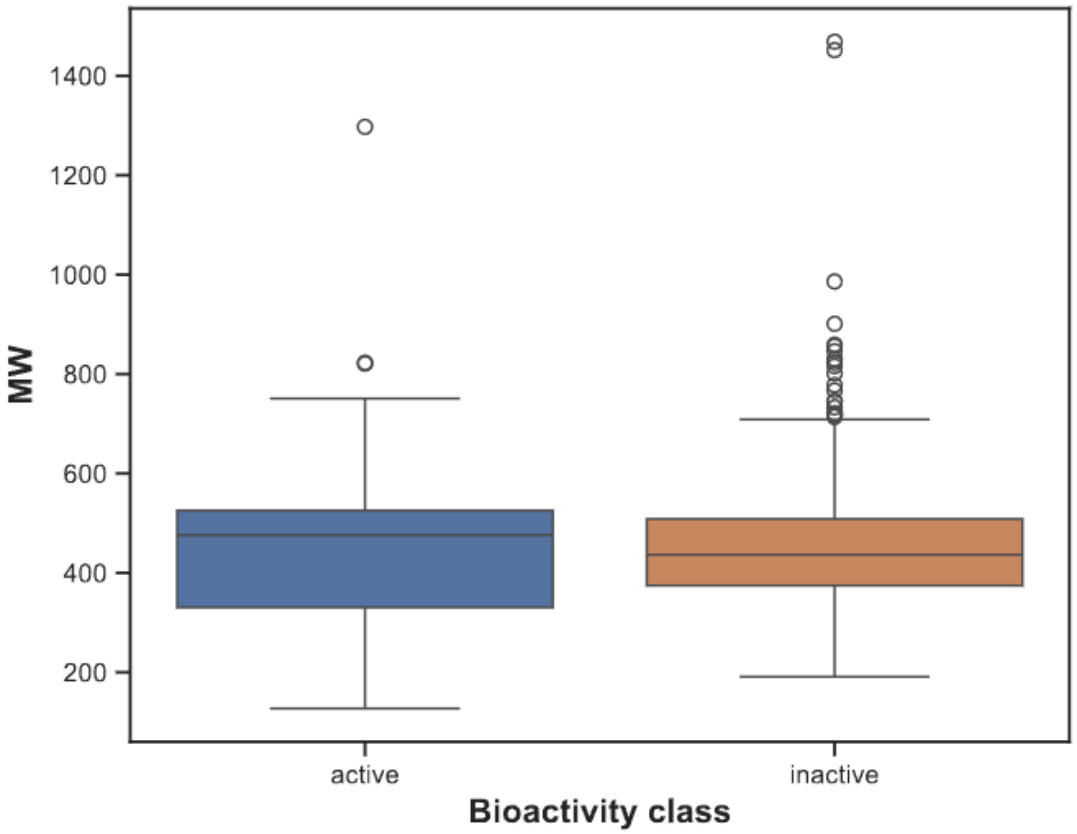
4.1 pIC50



Picture 3: Boxplot of pIC50 value by Bioactivity class.

Interpretation: Different distribution.The distribution of pIC50 values is different between active and inactive compounds. This is expected as different thresholds were set up to categorize different compounds as active or inactive.

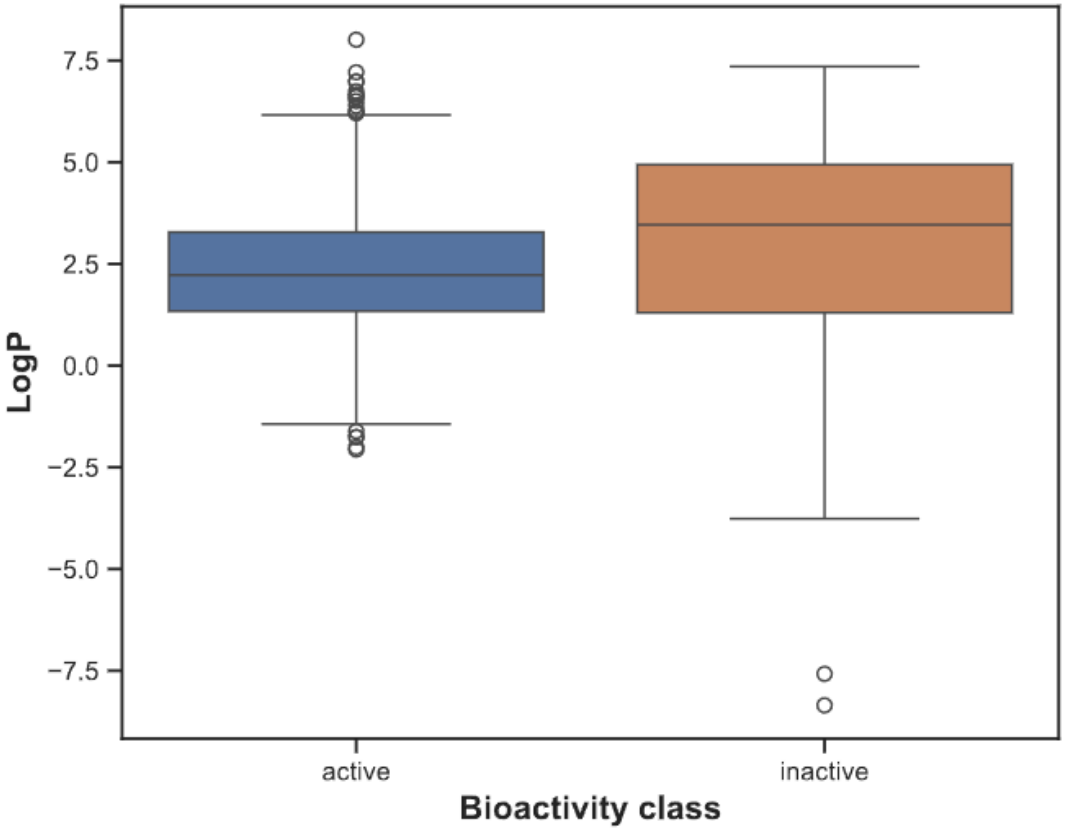
4.2 Molecular Weight



Picture 4: Boxplot of Molecular Weight by Bioactivity class.

Interpretation: Same distribution. The distribution of MW is similar between active and inactive compounds. This might indicate that the molecular weight of a molecule does not play a role in its bioactivity.

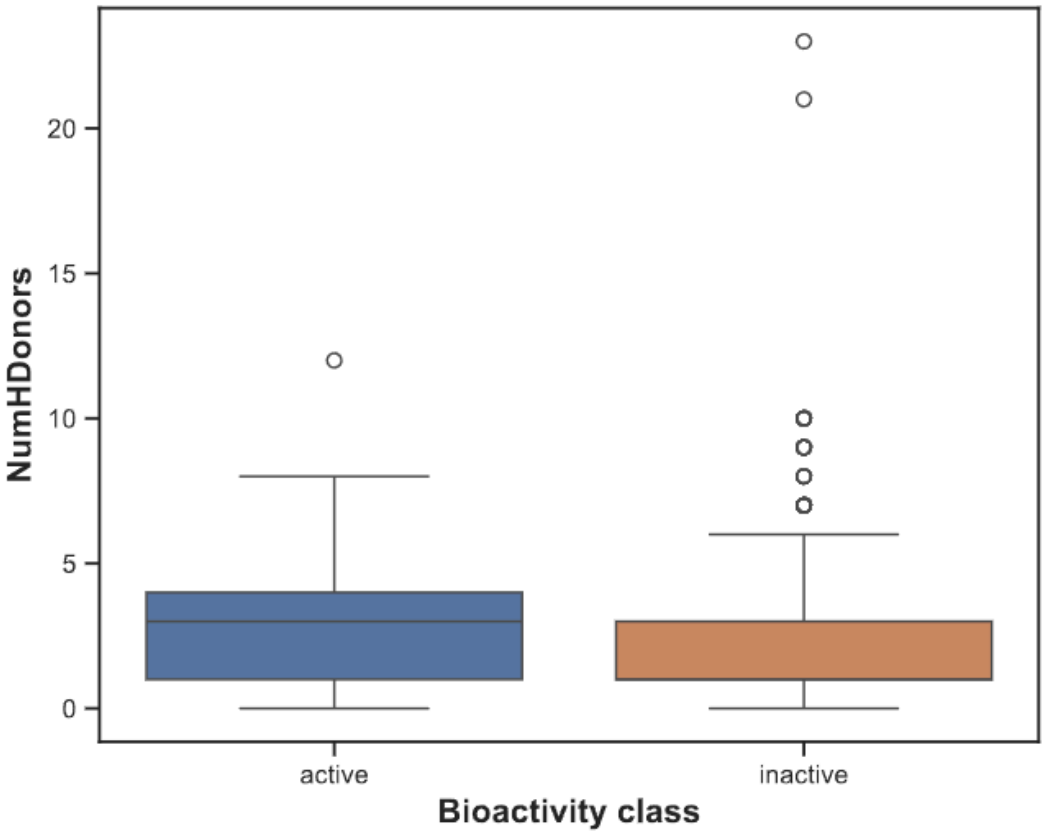
4.3 LogP



Picture 5: Boxplot of LogP by Bioactivity class.

Interpretation: Different distribution. The distribution of LogP is different between active and inactive compounds. This might indicate that the solubility of a molecule plays a role in its bioactivity.

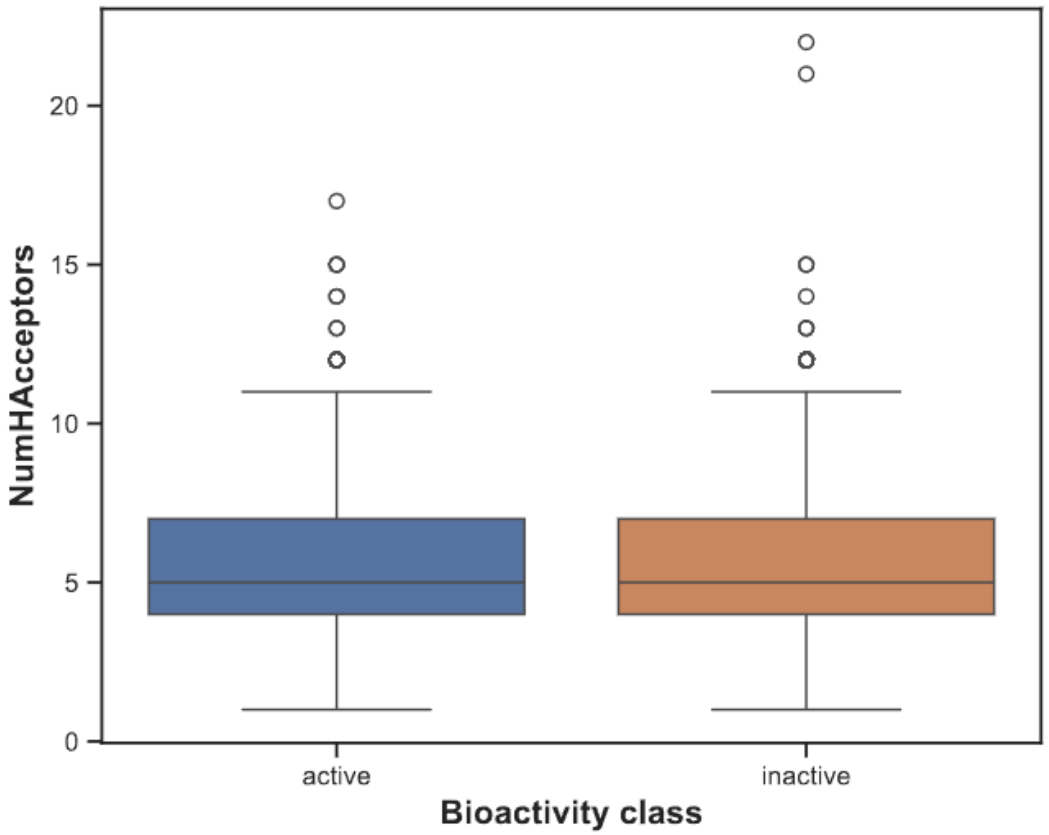
4.4 NumHDonors



Picture 6: Boxplot of NumHDonors by Bioactivity class.

Interpretation: Different distribution.The distribution of NumHDonors is different between active and inactive compounds. This might indicate that the number of hydrogen bond donors present in the molecule plays a role in its bioactivity.

4.5 NumHAcceptors



Picture 7: Boxplot of NumHAcceptors by Bioactivity class.

Interpretation: Same distribution.The distribution of NumHAcceptors is the same between active and inactive compounds. This might indicate that the number of hydrogen bond acceptors present in the molecule does not play a role in its bioactivity.

5. Regression Analysis

In this step we aimed to predict pIC50 values using Random Forest Regression. We created a dataset calculating the fingerprint descriptors of each molecule compound present in the previous dataset.

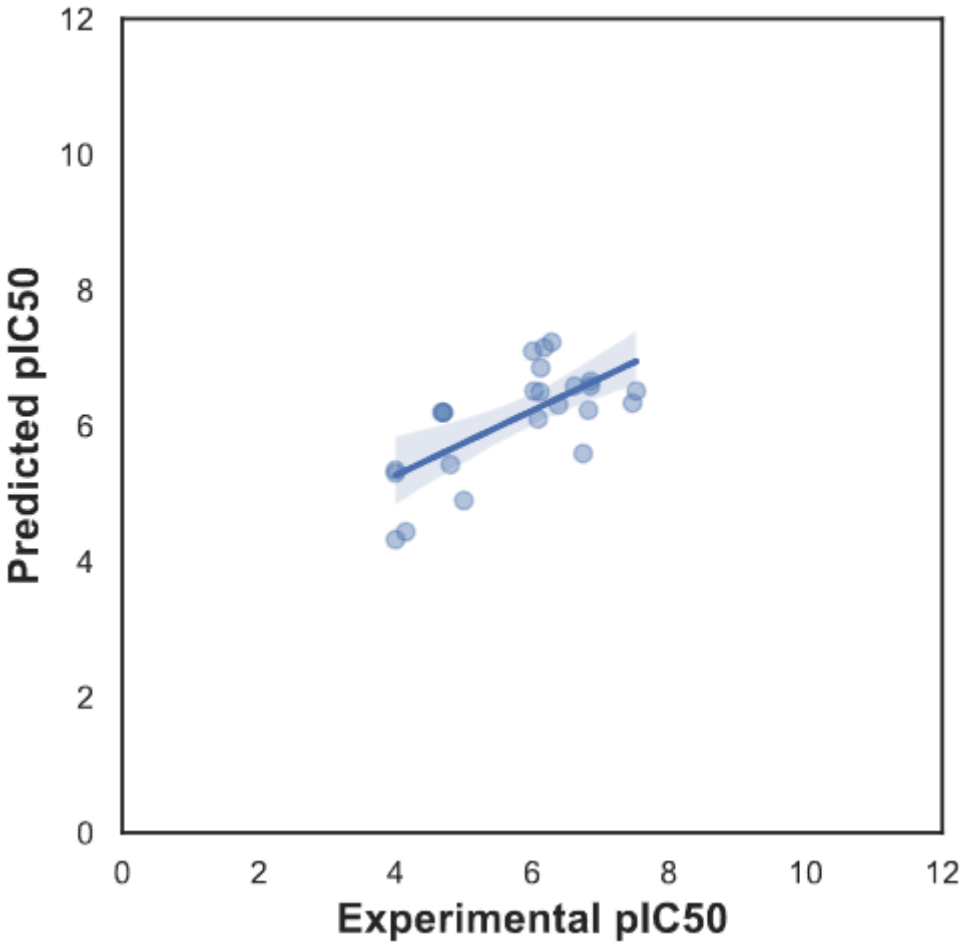
These fingerprint descriptors, most commonly known as **PubChem Fingerprints** (**PubChemFP),** are binary or count-based representations of molecular structures used in cheminformatics.They are generated from the chemical compounds in the PubChem database, a free resource for biological activities of small molecules.

PubChemFP uses a fingerprinting method to encode information about molecular structures. Each bit in a fingerprint corresponds to the presence or absence of specific substructures or features in the molecule.

Random Forest Regression, an ensemble learning method used for predicting continuous outcomes, was chosen for the prediction of the values. It operates by constructing multiple decision trees during training and outputting the average prediction of these trees for regression tasks.

For the training of the model the data was cleaned so as to not contain any incomplete or missing rows that could interfere with the performance of the model. The Random Forest regression model was trained on 80% of the dataset to predict the pIC50 values of compounds.

The model achieved a R^2 score of 0.3552760277195537.X on the test set. An R^2 score of **0.355** means that about **35.5%** of the variance in the test data is explained by the Random Forest regression model. This score suggests the model has some predictive power but is not particularly strong.

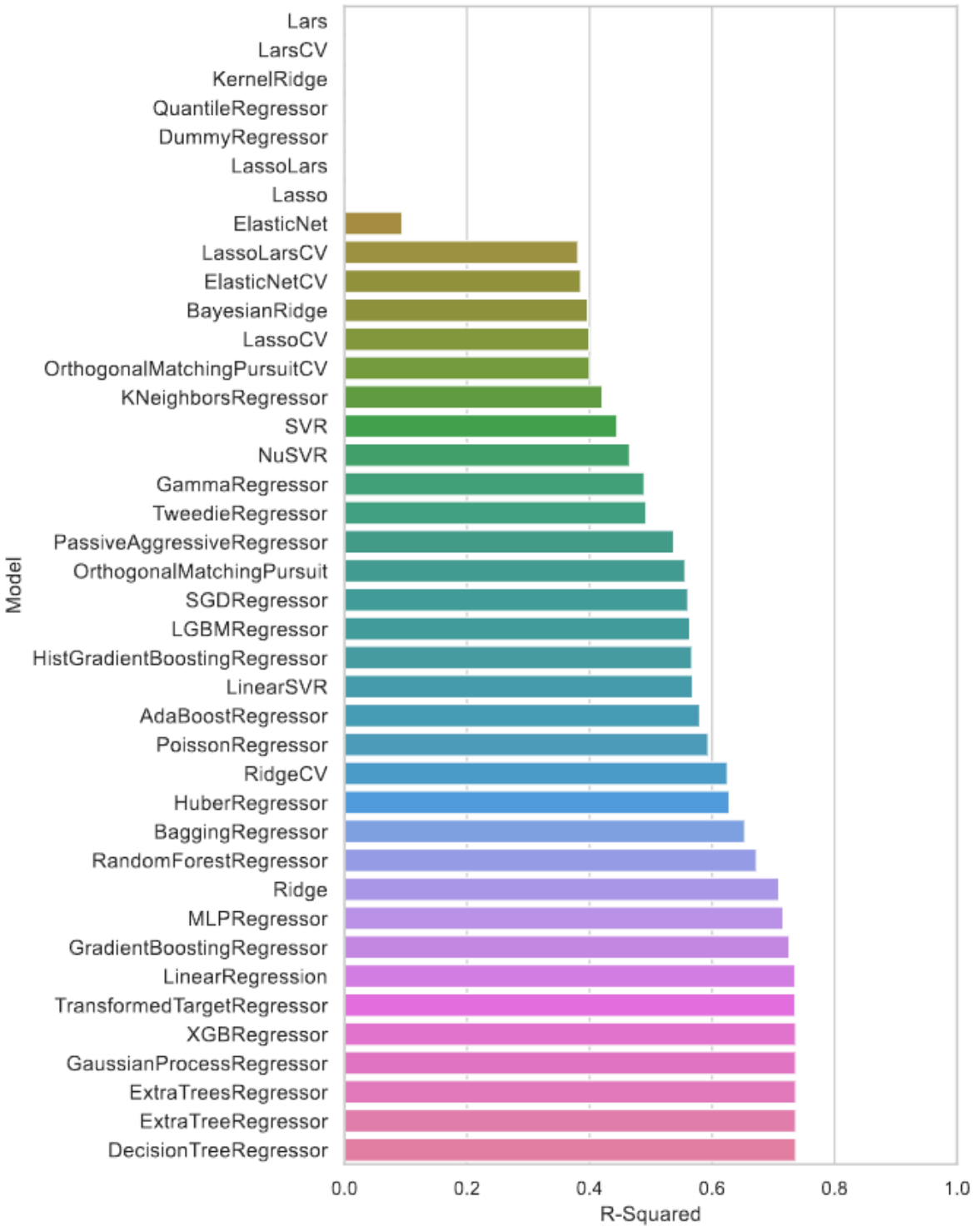


Picture 8: Scatter Plot of the regression model.

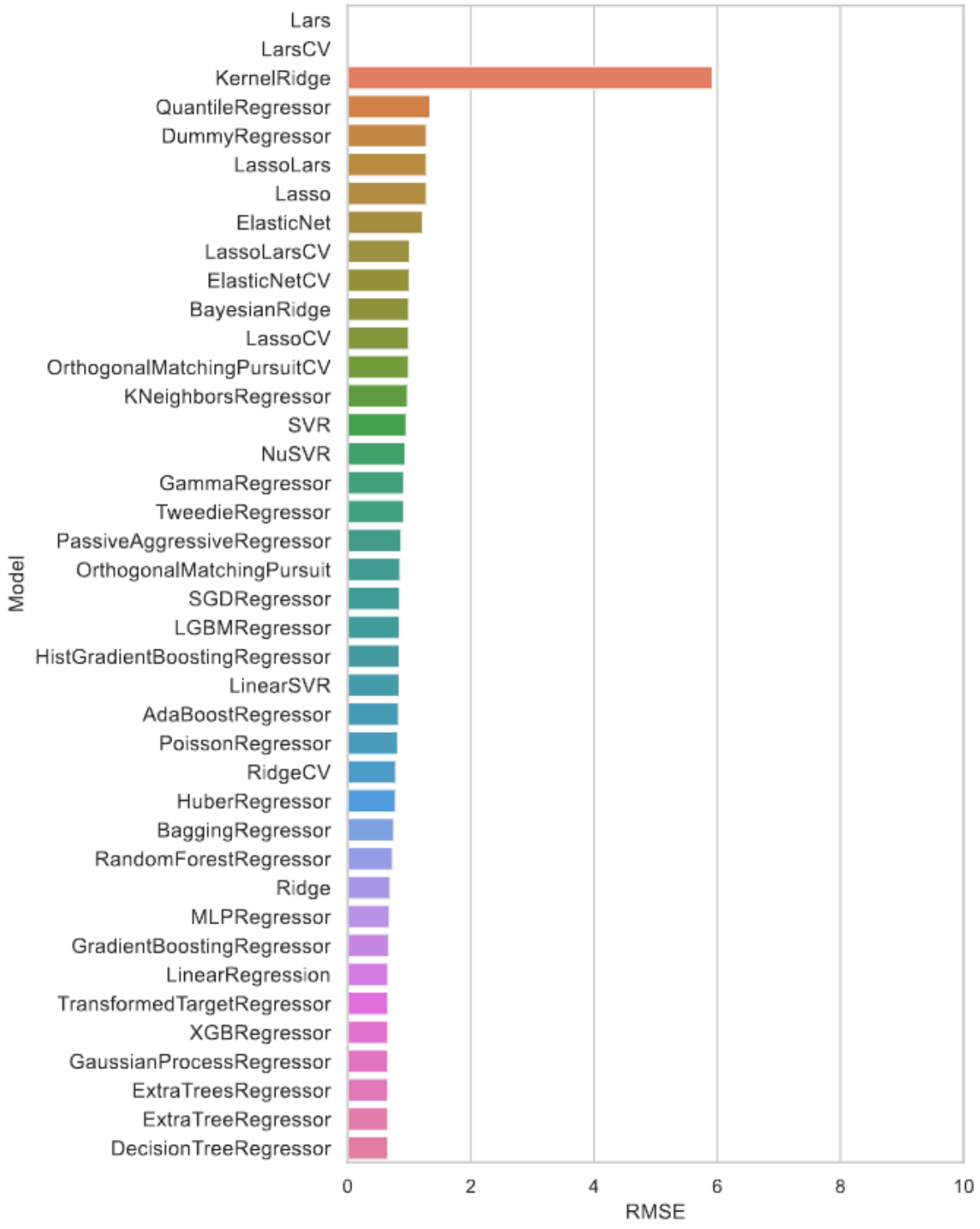
6. Model Comparison Using LazyPredict

We used the LazyRegressor to compare multiple machine learning models, analyzing calculation time, R^2 values and RMSE, a commonly used metric for evaluating the performance of regression models. RMSE is expressed in the same units as the target variable, making it easy to interpret. A lower RMSE indicates better model performance, while a higher RMSE indicates larger prediction errors. It measures the average magnitude of the prediction errors, giving insight into how well a model's predictions match the actual target values.

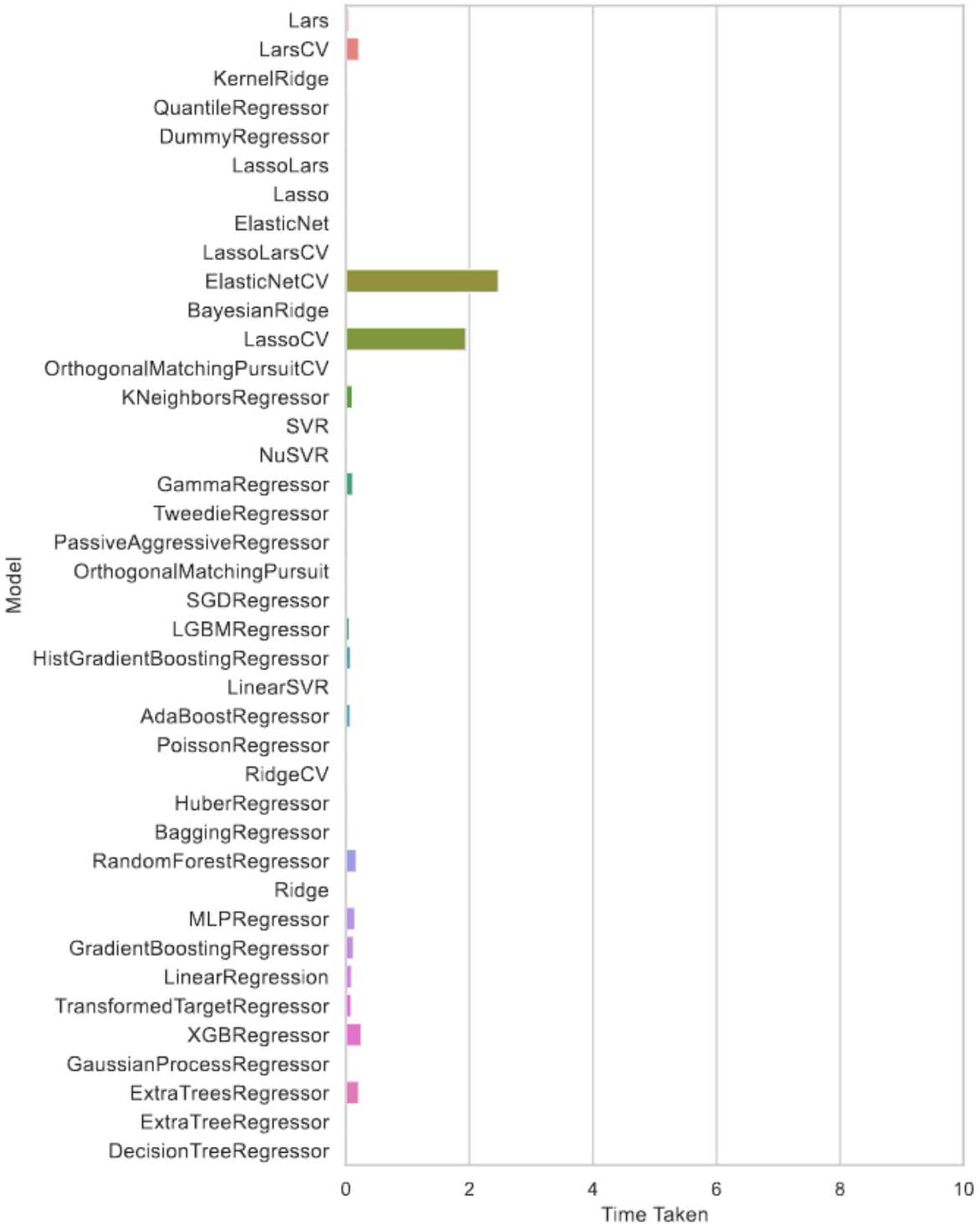
The top-performing models were DecisionTreeRegressor, ExtraTreeRegressor and GaussianProcessRegressor, all tied for first place with R^2 scores of 0.74, RMSE scores of 0.66 as well as the fastest calculation time, taking 0.01 seconds for the first two and 0.02 seconds for the GaussianProcessRegressor model.



Picture 9: Horizontal Bar plot displaying the R^2 values by model.



Picture 10: Horizontal Bar plot displaying the RMSE values by model.



Picture 11: Horizontal Bar plot displaying the calculation time, in seconds, by model.

7. Conclusion

The study successfully classified compounds based on their bioactivity, with significant differences in molecular descriptors between active and inactive compounds. Random Forest Regression provided strong predictions of pIC50, the DecisionTreeRegressor, ExtraTreeRegressor and GaussianProcessRegressor models emerged as the best performers in LazyPredict analysis.

Additionally, the code used for running this study as well as all of the data shown in this report can be found on:

https://github.com/FelipeASP2/Drug-discovery