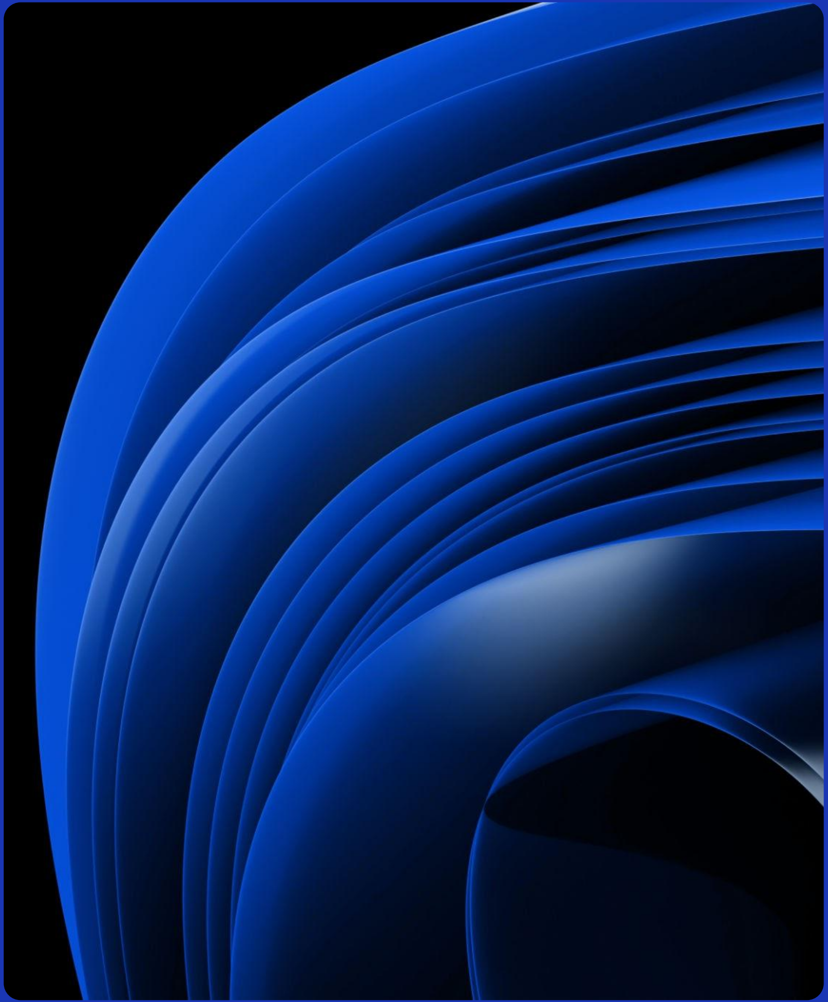
Predicting Kinase Selectivity Using Machine Learning

Confidential

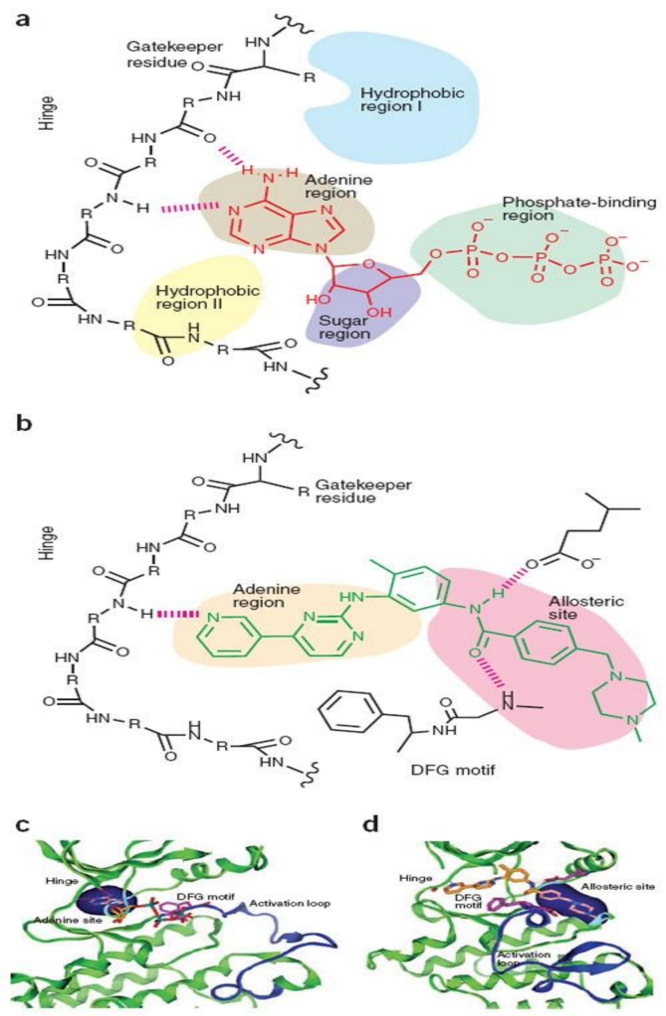
Copyright ©



Challenge

Use a machine learning model to rank kinases based on their likelihood of being inhibited by a specific compound

Use results to predict selectivity profiles for the specific inhibitors



Aspects of the Challenge

- Large volume of kinase - inhibitor combinations
- Many variables and data to consider - Significant effects on inhibitor binding
    - Kinase mutations
    - Kinase Families
    - Kinase amino sequences
    - Inhibitor binding modes
    - Inhibitor SMILES strings
- Importance of selectivity score
    - Low S (closer to 0%)
        - High selectivity, selective inhibitor
    - High S (closer to 100%)
        - Low selectivity, broader effects,, might cause side effects
- Developing an appropriate machine learning model to compute the data effectively and produce accurate predictions

Task Strategy Options

Regression

- Predicting continuous numerical values (K_d)
- Provides exact constants
- Flexibility in choosing own threshold
- Easy to generate new data points for unknown inhibitors and kinases
- Useful for ranking

Classification

- Less sensitive to outliers
- Less ml model feature tuning needed
- Easy to interpret
- Reduced sensitivity to minor errors and fluctuations in data

Possible Strategy

- Use both classification and regression to improve efficiency and accuracy

Classification

- Binary classification - Affinity and No Affinity
    - Significant percentage of 10001 K_d values, consider as No Affinity

Regression

- Apply regression to set classified as having an affinity
- Predict K_d values

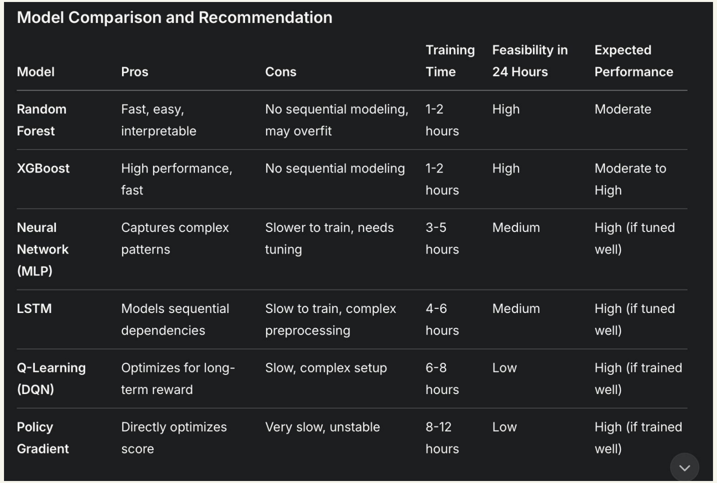
Model Selection

Model Options

- Random Forest
- XGBoost
- Neural Network
- LSTM
- Q-Learning
- Policy Gradient

Our Model - XGBoost

- High performance
- Fast
  - Low training time
- Feasible in given time
- Gradient Boosted Trees
- Useful and powerful for both classification and regression tasks



Model Comparison and Recommendation

Model	Pros	Cons	Training Time	Feasibility in 24 Hours	Expected Performance
Random Forest	Fast, easy, interpretable	No sequential modeling, may overfit	1-2 hours	High	Moderate
XGBoost	High performance, fast	No sequential modeling	1-2 hours	High	Moderate to High
Neural Network (MLP)	Captures complex patterns	Slower to train, needs tuning	3-5 hours	Medium	High (if tuned well)
LSTM	Models sequential dependencies	Slow to train, complex preprocessing	4-6 hours	Medium	High (if tuned well)
Q-Learning (DQN)	Optimizes for long-term reward	Slow, complex setup	6-8 hours	Low	High (if trained well)
Policy Gradient	Directly optimizes score	Very slow, unstable	8-12 hours	Low	High (if trained well)

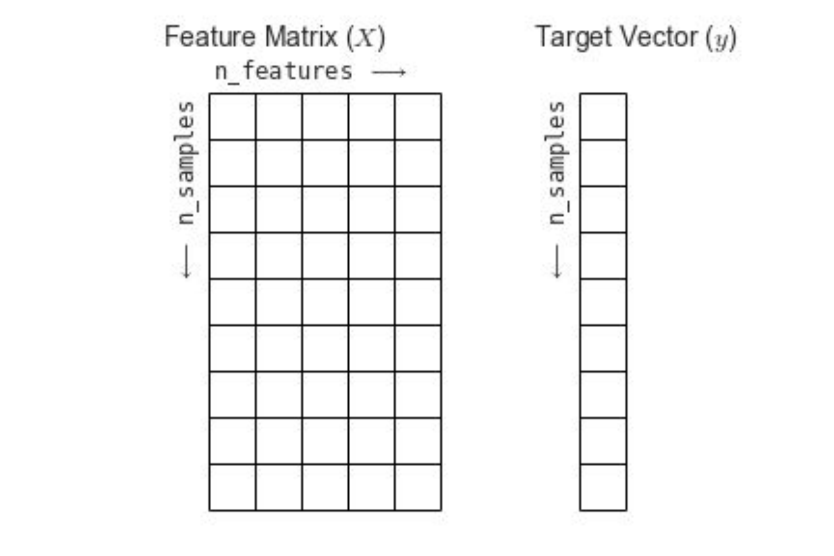
Feature Matrix

Goal:

- Create a feature matrix with one entry for each kinase-inhibitor combination
- Features for kinases and inhibitors

Strategy

- Identify most significant pieces of data and variables that affect the affinity of kinases and inhibitors - Set as features (input)
- Use provided dissociation constants as output for model training
- Desired predictions are K_d values



Data Processing

Kinase Group

Kinase group likely to display characteristics, and similar inhibitor affinities

Processing

- Label encoding

Kinase Amino Sequence

Protein sequence directly determines binding and dissociation constants

Processing

- ProteinBERT takes protein chain, applies deep learning, outputs NumPy array
- PCA reduces ProteinBERT array from 16000 features to 300, while preserving 100% variability

Kinase Mutations

Mutations can directly affect binding sites. Also allosteric changes, hydrophobicity, electrostatic change, etc.

Processing

- Label encoding

Data Processing

Inhibitor Binding Moes

Type 1 vs Type 2: Bind to ATP-binding site in active DFG-in conformation vs Bind to inactive DFG-out conformation of kinases. Affect affinity especially in conjunction with mutations.

Processing

- Label Encoding

Inhibitor SMILES strings

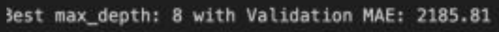
Chemical structure important for finding patterns that might affect affinity with kinases.

Processing

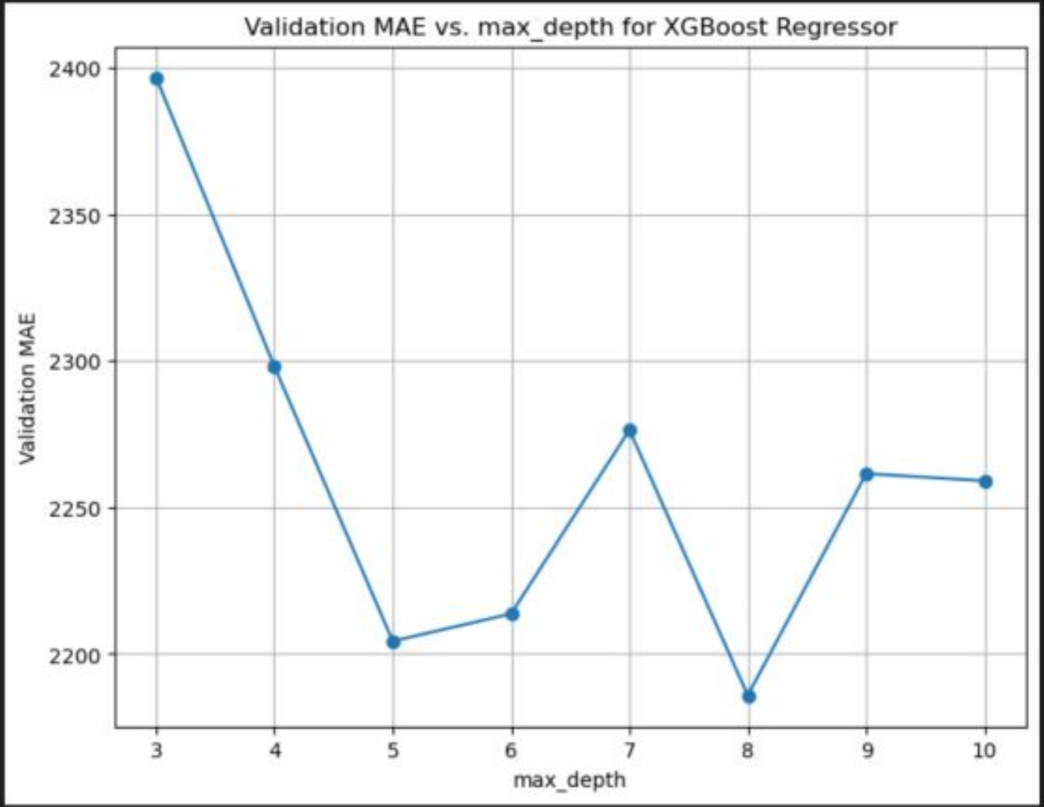
- RDKit takes SMILES strings as input and generaties corresponding molecular fingerprints that can be processed by a machine learning model

Results

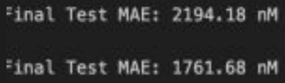
Regression



```
Best max_depth: 8 with Validation MAE: 2185.81
```



Evaluating the best depth for the tree on the Training data

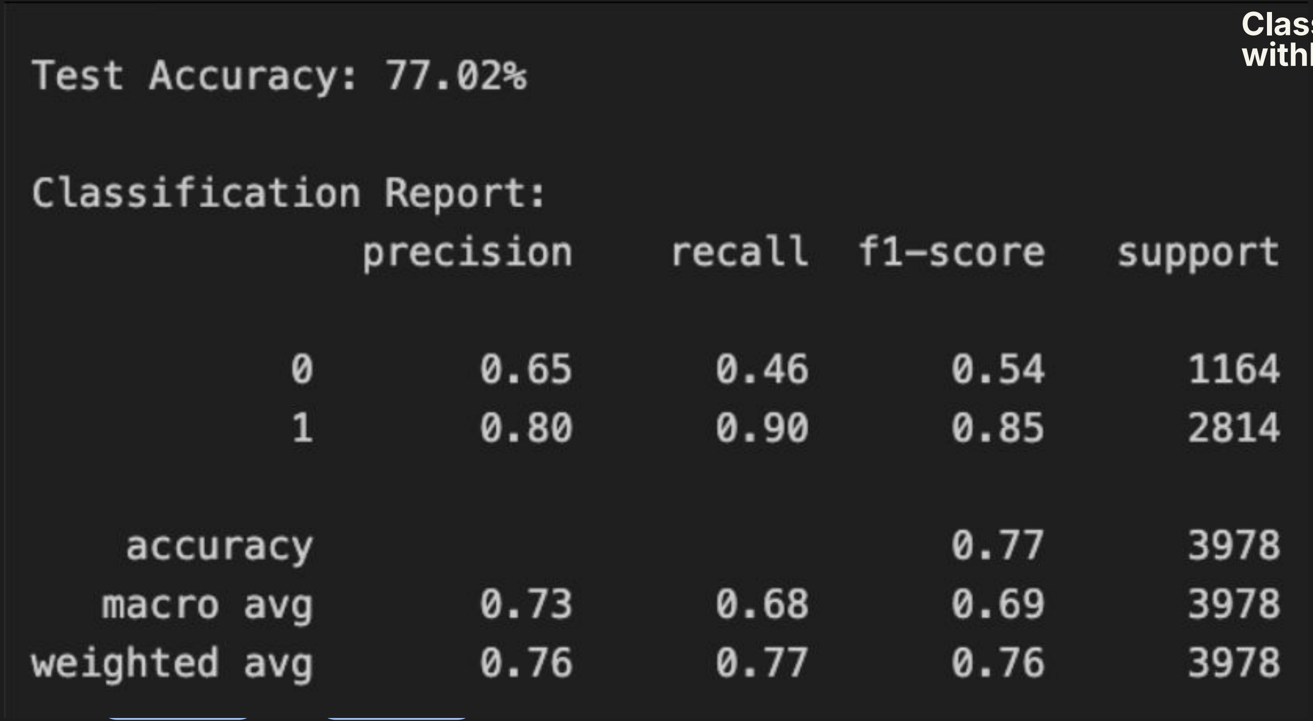


```
Final Test MAE: 2194.18 nM

Final Test MAE: 1761.68 nM
```

Confidential

Copyright ©

Classification

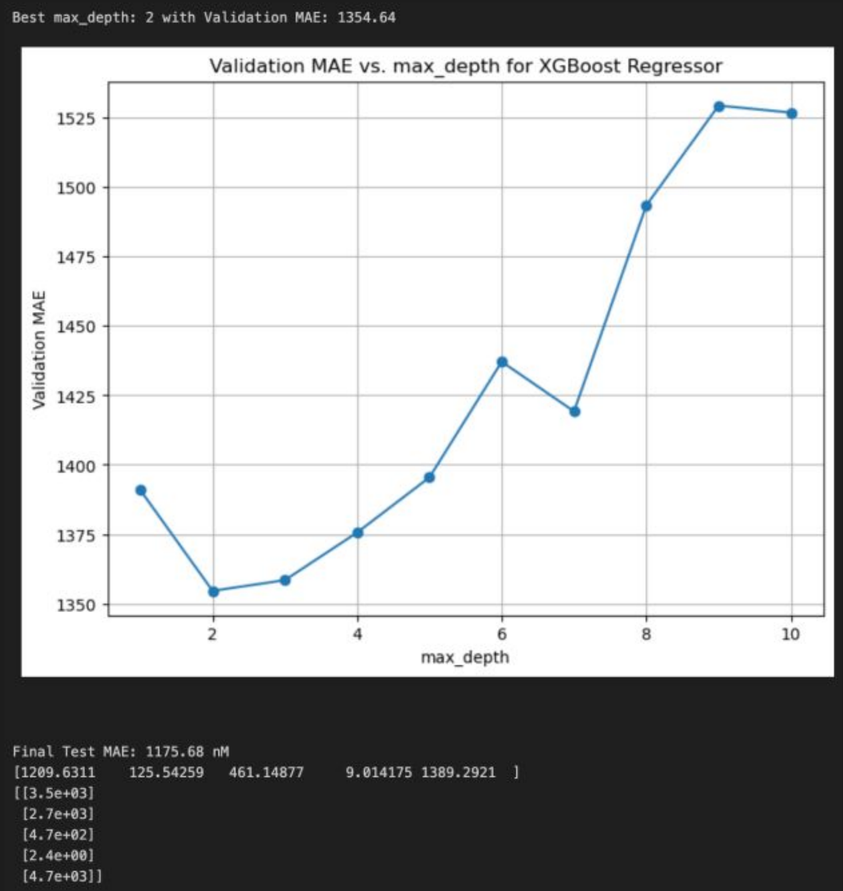


```
Test Accuracy: 77.02%

Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.46      0.54      1164
           1       0.80      0.90      0.85      2814

    accuracy                           0.77      3978
   macro avg       0.73      0.68      0.69      3978
weighted avg       0.76      0.77      0.76      3978
```

Classification on testing data withheld from the training set

Regression

Regression results after excluding data points of uncorrelated data

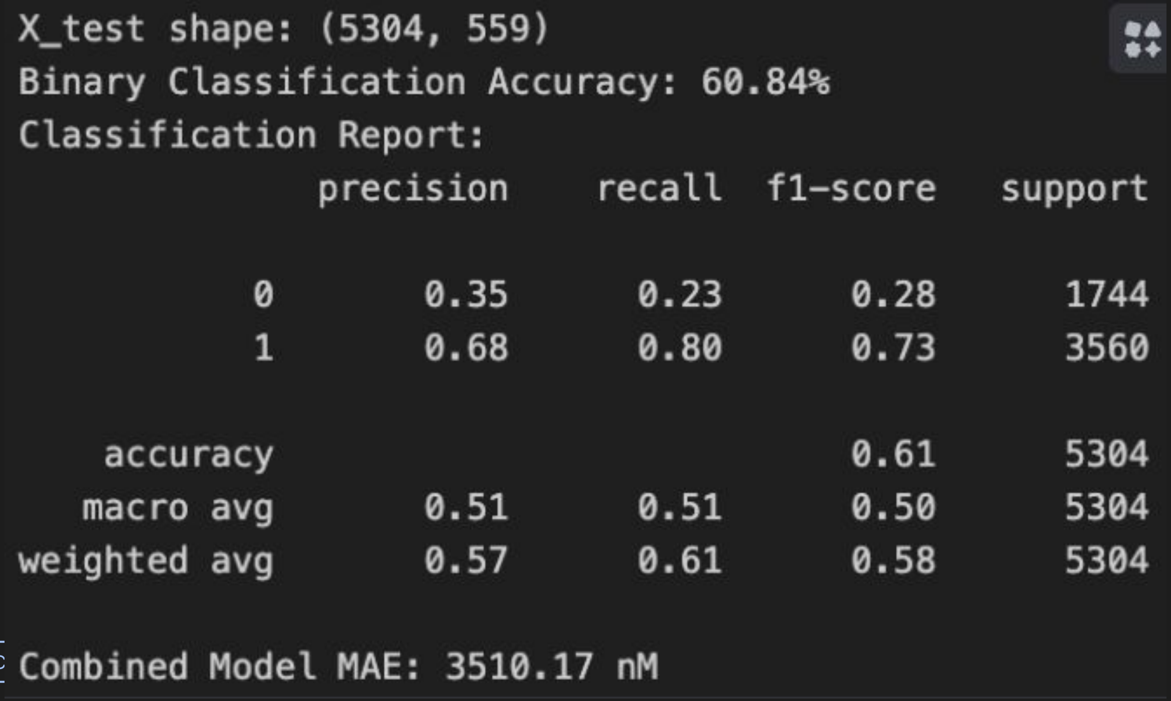


```
Best max_depth: 2 with Validation MAE: 1354.64
```

```
Final Test MAE: 1175.68 nM
[1209.6311    125.54259   461.14877     9.014175 1389.2921  ]
[[3.5e+03]
 [2.7e+03]
 [4.7e+02]
 [2.4e+00]
 [4.7e+03]]
```

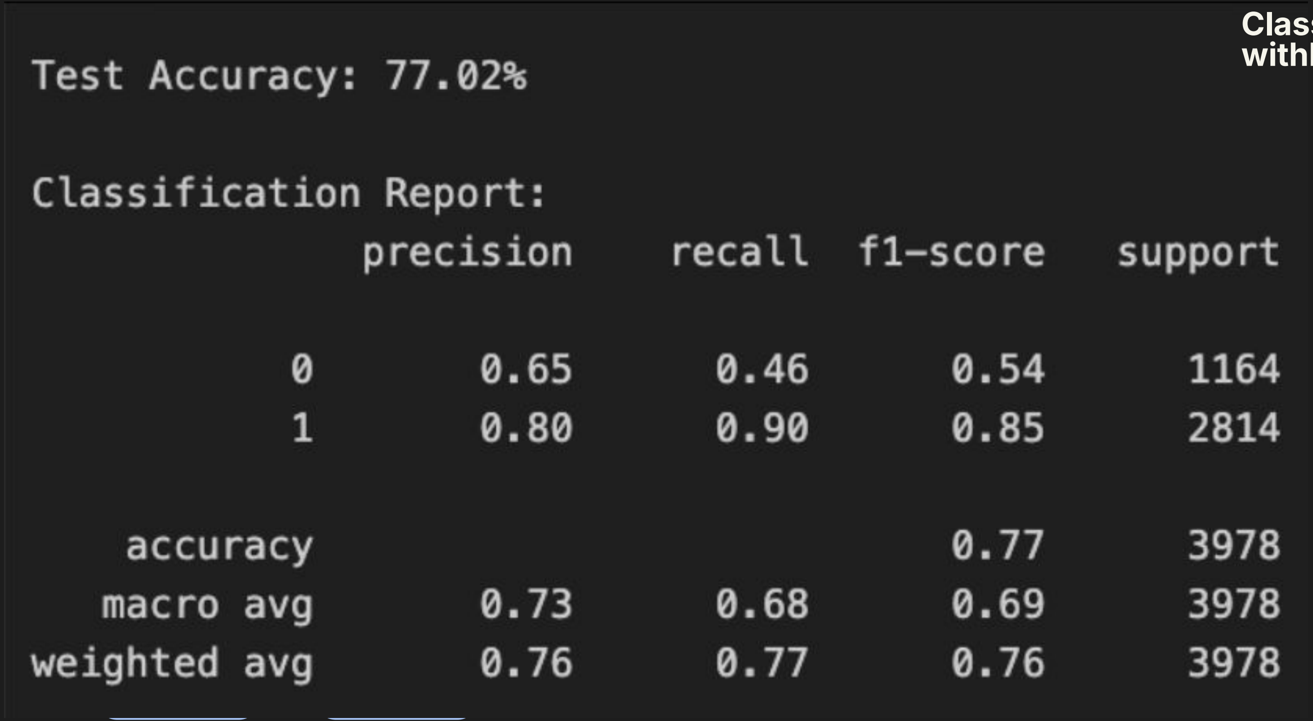
Confidential

Results on 12 withheld inhibitors



```
X_test shape: (5304, 559)
Binary Classification Accuracy: 60.84%
Classification Report:
              precision    recall  f1-score   support

           0       0.35      0.23      0.28      1744
           1       0.68      0.80      0.73      3560

    accuracy                           0.61      5304
   macro avg       0.51      0.51      0.50      5304
weighted avg       0.57      0.61      0.58      5304

Combined Model MAE: 3510.17 nM
```

Classification



```
Test Accuracy: 77.02%

Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.46      0.54      1164
           1       0.80      0.90      0.85      2814

    accuracy                           0.77      3978
   macro avg       0.73      0.68      0.69      3978
weighted avg       0.76      0.77      0.76      3978
```

Classification on testing data withheld from the training set

Result Validation

Evaluation Metrics

We decided to use MAE to calculate the K_d accuracy. As our end predictions are the exact K_d, MAE was the most appropriate method to evaluate our methods.

We decided to use accuracy, precision, recall, and F1 to evaluate accuracy of the classification model.

Conclusions

- Initially, our goal was to build a classification model that could predict whether a kinase-inhibitor interaction had a dissociation constant (Kd) below or above 3000 nM. However, after evaluating the model's performance, we found that it struggled to make accurate predictions in this range.
- Specifically, we treated compounds with a recorded Kd of **100001 nM** (which signifies no detectable binding) as one class and those with measurable affinity as the other. This binary approach significantly improved model performance, making it easier to distinguish between compounds that bind and those that do not (although the results were still not good)
- We noticed that the regression model that included all data performed better than the one that got rid of uncorrelated data