Kinase Selectivity Prediction

By The Kinines

Table of contents

01

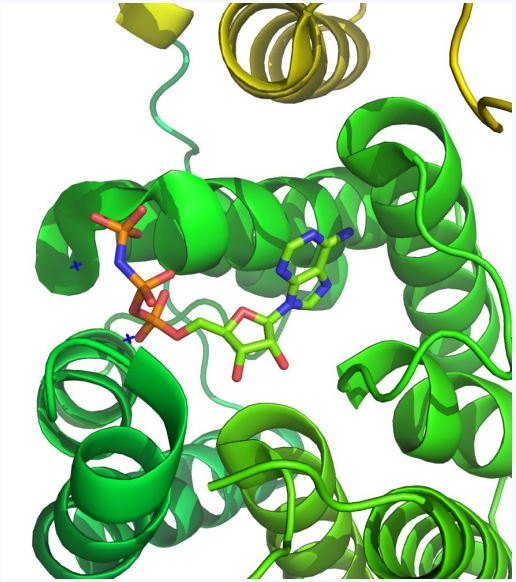
Problem Overview

Introduction

- Kinases regulate cellular functions and are key drug targets.
- Challenge: Predicting a kinase selectivity to minimize off-target effects.

Goal: Use **classification** to predict whether an inhibitor binds to a kinase at:

- **Kd < 300nM (strong binder)**
- **Kd < 3000nM (moderate binder)**



https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FKinase&psig=AOvVaw3qc3gwWUUHQpn83yolX7UJ&ust=1742227036093000&source=images&cd=vfe&opi=89978449&ved=0CBUQjRxqFwoTCKiMitD7jowDFQAAAAAdAAAAABAE

Dataset & Feature Engineering

Dataset Overview:

- 442 kinases, 60 inhibitors (48 train, 12 test)
- SMILES strings for compounds, kinase families, and selectivity scores

Explorations:

- Preprocessing: SMILES to molecular fingerprints (ECFP4), kinase embeddings (ProteinBERT)
- Data balancing: Handling class imbalance using class weights

02

Methodology

Strategy and Solution

Approach

- Separate binary classifications for 300 nM and 3000 nM thresholds.
- Using features:
  a. ECPF encoding for inhibitors (512 features)
  b. ProteinBERT embedding of the Kinases (PCA keeping 95% variance) (50k → ~100 features)

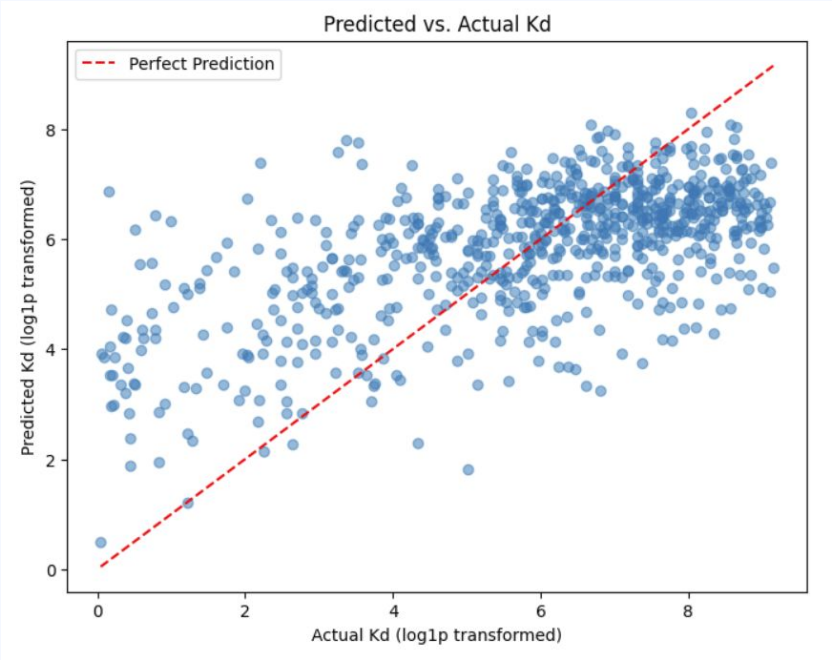
Tools

- RDKit for molecular fingerprints ⇒ capture compound structure
- XGBoost for classification ⇒ handles imbalanced data and provides interpretability.
- ProteinBERT for kinase embeddings ⇒ capture protein sequence information.

Attempts

Attempt #	Name	Details
1	XGBoost Regression (baseline)	ECPF4 fingerprints from RDKit and a XGB model
2	XGBoost Multiclass Classifier	ECPF4 fingerprints from RDKit and a XGB model
3	ECFP Sizes	Calibrated ECFP bit length to about 1/10 of example count
4	SMOTE/SMOTETomek	Applied SMOTE and Tomek to balance under/over-represented data
5	XGBoost Logistic Classifier	Separated classification into two binary classification tasks
6	scale_pos_weight	Add balance parameter to XGBoost Logistic Classifier
7	Threshold calibration	Calibrate threshold for prediction probabilities

Baseline model



Mean Squared Error: 3.3050043833939533
Mean Absolute Error: 1.4468829188735501
Mean R^2 Error: 0.35691339345649853

Fingerprints only.

Identifying other features

S (300nM)

S(300nM)	1.000000e+00
S(3000nM)	9.301859e-01
Binding Mode (based on ABL1-phos. vs. -nonphos affinity)_Type I	3.808686e-01
Mutant_NO	1.523698e-15

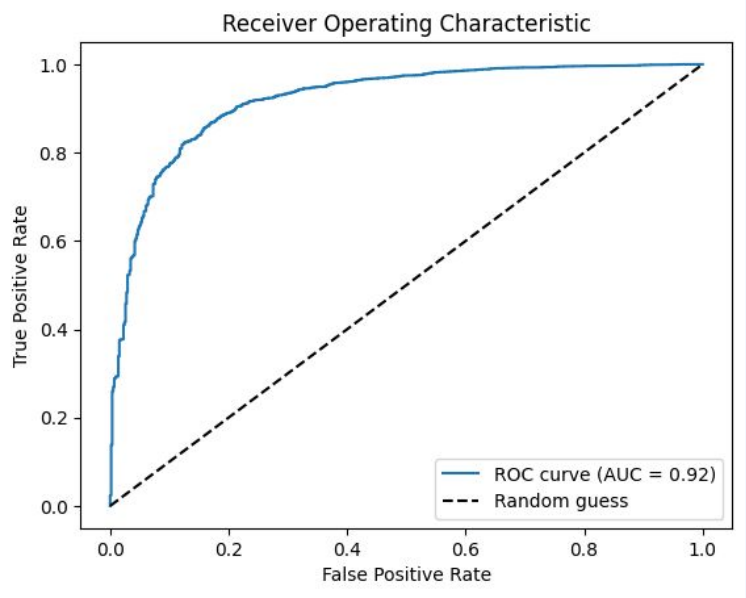
S (3000nM)

S(3000nM)	1.000000e+00
S(300nM)	9.301859e-01
Binding Mode (based on ABL1-phos. vs. -nonphos affinity)_Type I	5.038162e-01
Mutant_NO	1.849854e-15

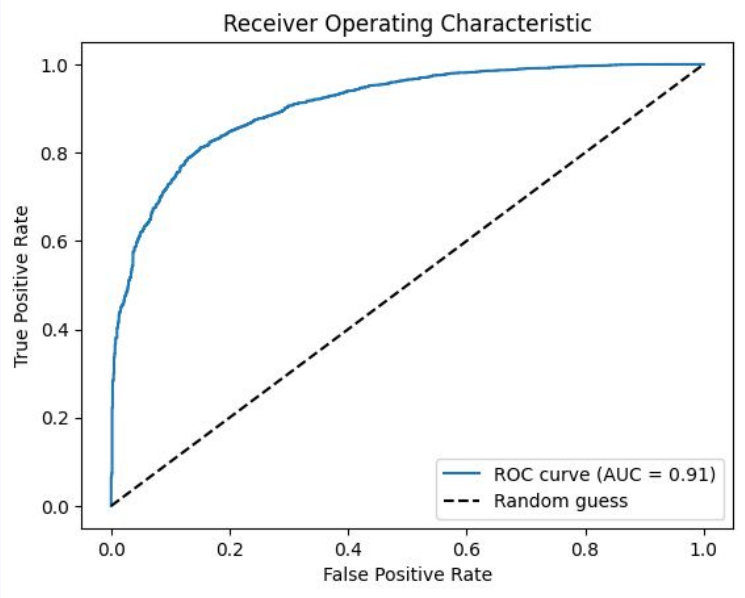
Didn’t find significant correlations for other features. Tried with binding mode, not a significant improvement over just ProteinBERT embeddings + fingerprints. May have benefited from better exploration of features, or finding some way to incorporate the binding mode more heavily.

ROC

S(300nM)



S(3000nM)



Model Training & Evaluation

Classification models

- Separate models for 300 nM and 3000 nM thresholds
- Metrics: Accuracy, Precision, Recall, F1-Score, MSE, Zero-One Loss, Selectivity Score

Validation

- Train-test split (80-20)
- Stratified sampling to handle class imbalance.

⇒ Adjusted for imbalanced data using `scale_pos_weight`

Model Training & Evaluation

Parameters

```
params = {
        'objective': 'binary:logistic',
        'eval_metric': 'logloss',
        'scale_pos_weight': scale_pos_weight,
        'max_depth': 3,
        'eta': 0.1,
        'subsample': 0.8,
        'colsample_bytree': 0.8,
}
```

Experimented with:

- Learning rate (best results at 600 steps)

Attempted hyperparameter optimization with GridSearchCV, but took too long to run.

03

Results

Results - 300nM

Model	Accuracy	F1-Score	Precision	Recall
<300 nM	0.9013	0.62	0.54	0.72
>300 nM		0.94	0.96	0.92

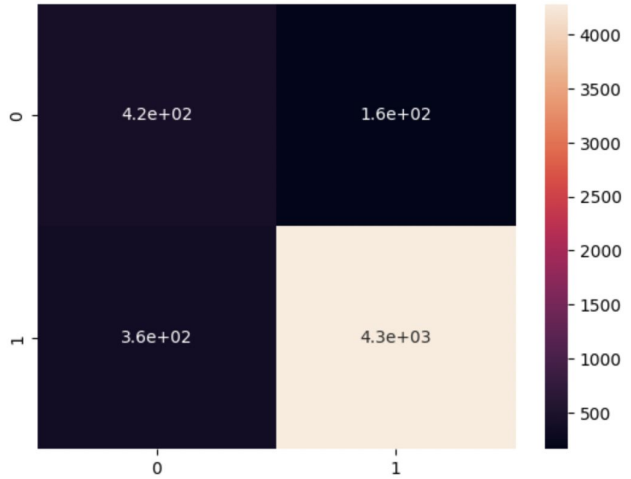
Zero-One Loss: 0.0987

```
Accuracy: 0.9013
              precision    recall  f1-score   support

      <300nM       0.54      0.72      0.62       578
      >300nM       0.96      0.92      0.94      4638

    accuracy                           0.90      5216
   macro avg       0.75      0.82      0.78      5216
weighted avg       0.92      0.90      0.91      5216

Zero-One Loss: 0.0987
```



Results - 3000nM

Model	Accuracy	F1-Score	Precision	Recall
<3000 nM	0.8545	0.70	0.68	0.72
>3000 nM		0.90	0.91	0.90

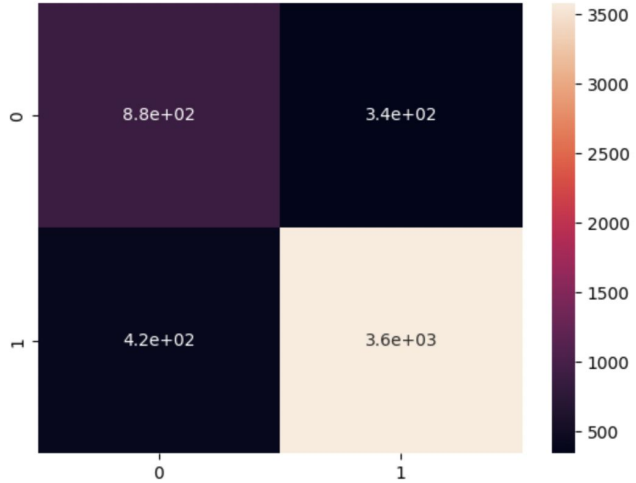
Zero-One Loss: 0.1455

```
Accuracy: 0.8545
              precision    recall  f1-score   support

     <3000nM       0.68      0.72      0.70      1221
     >3000nM       0.91      0.90      0.90      3995

    accuracy                           0.85      5216
   macro avg       0.80      0.81      0.80      5216
weighted avg       0.86      0.85      0.86      5216

Zero-One Loss: 0.1455
```



Validation and Blind data

Model	Accuracy	F1-Score	Precision	Recall
300 nM	0.8740	0.84	0.84	0.87
3000 nM	0.7304	0.72	0.72	0.73

Model	MSE	Zero-One Loss	Mean Selectivity Score (actual)
300 nM	0.0251	0.125	0.111
3000 nM	0.0483	0.270	0.250

300nM

	Compound	S(300nM) predicted	S(300nM)
0	AMG-706	0.083710	0.0389
1	BIBF-1120 (derivative)	0.067873	0.2927
2	CI-1040	0.176471	0.0026
3	GSK-461364A	0.305430	0.0155
4	PI-103	0.045249	0.0207
5	SKI-606	0.223982	0.1917
6	Sorafenib	0.149321	0.0803
7	SU-14813	0.061086	0.2124
8	Sunitinib	0.018100	0.3109
9	TG-100-115	0.147059	0.0337
10	VX-680/MK-0457	0.210407	0.1321
11	VX-745	0.033937	0.0052

3000nM

	Compound	S(3000nM) predicted	S(3000nM)
0	AMG-706	0.081448	0.0777
1	BIBF-1120 (derivative)	0.210407	0.5181
2	CI-1040	0.117647	0.0078
3	GSK-461364A	0.423077	0.1010
4	PI-103	0.156109	0.0570
5	SKI-606	0.414027	0.4249
6	Sorafenib	0.248869	0.1684
7	SU-14813	0.264706	0.5415
8	Sunitinib	0.074661	0.5959
9	TG-100-115	0.171946	0.1321
10	VX-680/MK-0457	0.346154	0.3472
11	VX-745	0.081448	0.0233

04

Conclusion

Key Findings

Unbalanced Dataset

The greatest challenge was the extreme imbalance in the dataset

Binary predictions

Significant improvement in accuracy after switching to binary predictions

Recall vs. Precision

Some models performed better on recall, while other performed better on precision

Future Research

Multiple Sequence Alignment for Kinase Selectivity Prediction

- **Identify Commonalities Across Kinase Families:**
  - Perform multiple sequence alignment (MSA) to uncover conserved patterns in:
    - **Active sites** – Key residues involved in catalytic activity.
    - **Binding sites** – Regions critical for substrate and inhibitor interactions.
    - **Regulatory motifs** – Sequences involved in allosteric regulation.

- **Leverage Embeddings from Common Sequences:**
  - Generate embeddings from aligned sequences to capture evolutionary and functional relationships.
  - Integrate these embeddings into predictive models to:
    - Improve kinase-inhibitor binding specificity predictions.
    - Enhance generalization across kinase families.
    - Uncover novel selectivity patterns and off-target effects.

Thanks !

Any questions?