CHALLENGE: PREDICT KINASE SELECTIVITY USING MACHINE LEARNING

PharmaHacks 2025
contact@pharmahacks.com

Molecular Forecaster
agathe.fayet@molecularforecaster.com
ben.weiser@molecularforecaster.com

ABSTRACT

Kinases are a large family of enzymes that regulate various cellular activities. Predicting kinase selectivity using machine learning can accelerate drug discovery by prioritizing compounds for experimental validation. This challenge provides datasets and asks participants to develop models for predicting kinase selectivity profiles of small molecule inhibitors.

1 INTRODUCTION

Kinases play a crucial role in cellular processes, making them important drug targets in oncology and other diseases. However, achieving selectivity is challenging, as inhibitors often bind to multiple kinases, leading to off-target effects.

Davis et al. (2011) studied interactions between 72 kinase inhibitors and 442 kinases, covering over 80% of the human protein kinome. These kinases include mutants and activation variants, resulting in 386 distinct kinase domains. Kinases within the same family share structural similarity, explaining why inhibitors targeting one kinase may bind to others in the same family.

2 KINASE FAMILIES

The dataset includes kinases from several families:

- Tyrosine Kinase (TK)
- Tyrosine Kinase-Like kinases (TKL)
- Protein Kinase A, G, and C families (AGC)
- CDK, MAPK, GSK3, and CLK families (CMGC)
- Calcium/calmodulin-dependent Protein Kinase (CAMK)
- Homologs of yeast Sterile Kinases (STE)
- Other kinases
- Atypical Kinases
- Lipid Kinases
- Kinases from human pathogens

3 EXPERIMENTAL DATA

For each inhibitor-kinase, the equilibrium dissociation constant (K_d) in nanomolar (nM) was measured for inhibitor-kinase interactions. A lower K_d indicates a higher affinity. Table 1 presents sample data.

Accession Number	Kinase/Inhibitor couple	K_d (nM)
NP_055726.3	AAK1/A-674563	43
NP_005148.2	ABL1(E255K)-phosphorylated/AB-1010	140

Table 1: Sample inhibitor-kinase dissociation constant values.

4 SELECTIVITY SCORE

Davis et al. introduced the following selectivity score:

$$S_{\text{inhibitor}}(3000 \text{ nM}) = \frac{n_{\text{kinase}}(K_d < 3000 \text{ nM})}{386} \tag{1}$$

where:

- $n_{\text{kinase}}(K_d < 3000 \text{ nM})$ is the number of kinases bound by the inhibitor with high affinity ($K_d < 3000$ nM).
- 386 is the number of distinct kinase domains.

Table 2 presents selectivity scores.

Compound	Binding Mode	$S(300 \text{ nM})$	$S(3000 \text{ nM})$
A-674563	Undetermined	0.1166	0.2772
AB-1010	Type II	0.0337	0.0622

Table 2: Selectivity scores for kinase inhibitors.

Table 2 also contains SMILES strings for each compound. SMILES (Simplified Molecular Input Line Entry System) is a notation that represents chemical structures using ASCII strings, facilitating computational processing of molecular information.

5 CHALLENGE

Participants will develop a machine learning model to predict kinase selectivity profiles of small molecule inhibitors. The goal is to classify or rank kinases based on their likelihood of being inhibited by a given compound. Participants can either try to predict the **exact** equilibrium dissociation constant, which would make it a **regression** task, while participants can alternatively predict if K_d satisfies a certain threshold, i.e. $K_d < 300nM$ or $K_d < 3000nM$, in which case participants would perform a **classification** task. For both approaches, the participants will then **calculate the selectivity scores** using their earlier predictions and the given formula.

5.1 DATASETS

Participants will have access to:

- A list of 442 kinases (Table 1).
- A list of 60 inhibitors with SMILES strings and selectivity scores (Table 2). Twelve inhibitors will be removed for testing and provided separately.
- Experimental data on inhibitor-kinase binding, including dissociation constants (K_d).

5.2 EVALUATION METRICS

To evaluate your submissions, you can use the following evaluation metrics, although **you're free to use other ones**:

5.2.1 REGRESSION APPROACH

- **MSE** loss to evaluate S accuracy : $MSE = \frac{1}{n}\sum_{i=1}^{n}(S_i - \hat{S}_i)^2$
- **MAE/L1** loss to evaluate K_d accuracy : $MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{K_d} - K_d|$

5.2.2 CLASSIFICATION APPROACH

- **MSE** loss to evaluate S accuracy : $MSE = \frac{1}{n}\sum_{i=1}^{n}(S_i - \hat{S}_i)^2$
- **Zero-One Loss** to evaluate classification accuracy (for both 300nM and 3000nM thresholds) : $L(K_{d_{\text{pred}}}, K_d) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(K_{d_{\text{pred}}}^{(i)} < 3000 \neq {K_d}^{(i)} < 3000\right)$

5.3 PRESENTATION

You are expected to develop a presentation showcasing your explorations, results, and conclusions from the hackathon challenge. We understand that the timeframe is limited and that you may not have the opportunity to explore every aspect in depth or train on the full dataset. To accommodate this, we will use the following evaluation schema:

1. Explorations (15 points) Your team will receive a score from 1 to 15 points, primarily based on the following criteria:

- **Understanding of the Problem**: How well your team identifies the key components of kinase selectivity prediction, including the challenges of off-target effects, kinase families, and the importance of selectivity scores.
- **Strategy and Solution**: How clearly your team defines a strategy to predict kinase selectivity, whether through regression or classification, and how you plan to handle the dataset (e.g., using SMILES strings, molecular fingerprints, or protein embeddings).
- **Explanation and Justification**: How effectively your team communicates the chosen approach, ensuring it is clear, complete, and well-justified. This includes explaining why specific machine learning models, tools (e.g., RDKit, ChemBERTa, ProteinBERT), or evaluation metrics (e.g., MSE, MAE, Zero-One Loss) were selected.

2. Results (25 points) Your team will receive a score from 1 to 25 points, primarily based on the following criteria:

- **Performance on Blind Data**: How well your approach predicts kinase selectivity on unseen data (e.g., the 12 withheld inhibitors) and optimizes the selectivity score using the provided formula.
- **Quality of Solutions**: The ability of your approach to identify effective solutions, even if they are not always optimal. This includes how well your model predicts *Kd* values or classifies inhibitors based on the given thresholds (300 nM or 3000 nM).
- **Validation Strategy**: The implementation of a structured strategy to ensure that all problem requirements are met, including the calculation of selectivity scores and the use of appropriate evaluation metrics (e.g., MSE, MAE, Zero-One Loss).

3. Conclusions (10 points) Your team will receive a score from 1 to 10 points, primarily based on the following criteria:

- **Presentation Quality**: Whether the presentation includes all necessary, relevant, and expected components, such as an overview of the problem, methodology, results, and conclusions.
- **Clarity and Innovation**: How well the proposed approach is articulated, ensuring it is innovative, clear, unambiguous, and quantified. Highlight any unique or creative aspects of your solution.
- **Engagement and Communication**: The effectiveness of visual elements (e.g., charts, graphs, molecular diagrams) and language in enhancing audience understanding and engagement. Ensure that your presentation is concise and well-structured

5.4 NOTES AND HINTS

You can quickly construct a baseline using only the compound-only method, employing ECPF4 fingerprints from RDKit and a XGB model. This baseline will serve as your target to surpass and benchmark your future iterations. However, it's a challenging baseline to achieve.

Avoid considering the entire protein in your models. This approach will likely result in an input dimensional space that's too vast for effective training on a dataset as small as the one used in this challenge. Additionally, there may be multiple binding sites per kinase, so there's no assurance that a compound in question binds to a specific location on the protein.

6 RESOURCES

The following is a list of resources that can be helpful. Using them or not will not affect your final evaluation:

- **ChemBERTa**: A transformer-based language model tailored for molecular property prediction. The hidden states produced by ChemBERTa can serve as embeddings, capturing intricate chemical information that enhances the accuracy of molecular property predictions.
- **RDKit** : An open-source software suite for cheminformatics, computational chemistry, and predictive modeling. It utilizes SMILES (Simplified Molecular Input Line Entry System) strings to generate molecular fingerprints, which are essential for tasks like molecular similarity assessments and drug discovery.
- **ProteinBERT**: A deep learning model that applies transformer architectures to protein sequences. The hidden states generated by ProteinBERT can function as embeddings, capturing detailed information about amino acid sequences, thereby improving the prediction of protein functions and interactions.
- **NCIB Protein Database** : The Protein database is a collection of amino acid sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB.

7 DISCLAIMER

The test set will be released at 11:00 a.m. on Sunday.

For submission, put your notebook and slides on a GitHub repository and post it on Devpost.