# **Project Proposal: Healthcare Analytics for Drug Discovery and Development**

Group 9:

Mitali Selot

Nandivardhan Reddy Bhumireddy

Sai Phani Sudheer Korlapati

**1. Problem Statement**: **Drug-Target Interaction (DTI) Prediction**

Drug discovery is a complex and resource-intensive process that involves identifying potential drugs and understanding their interactions with biological targets (proteins). Drug-Target Interaction (DTI) prediction is a crucial step in this pipeline, as it helps **prioritize drug candidates** before costly laboratory experiments and clinical trials. Current methods for DTI prediction rely on **high-throughput screening (HTS) and computational docking,** which are expensive, time-consuming, and often produce a high rate of false positives. With advancements in **machine learning (ML) and deep learning (DL), computational approaches offer a more efficient and scalable way to predict potential drug-target interactions.**

**2. Objectives**

The goal of this project is to develop a multi-label classification model for predicting Drug-Target Interactions (DTI) using advanced machine learning and deep learning techniques. Instead of simply predicting binding affinity, our model will classify drug-target pairs into **four distinct** interaction types based on their biological effects:

* **Activator** – The drug binds to the target and enhances its biological activity.
* **Inhibitor** – The drug binds to the target and reduces or blocks its activity.
* **Partial Agonist** – The drug binds to the target and activates it, but only partially, compared to a full activator.
* **Antagonist** – The drug binds to the target and prevents its activation by other molecules, effectively neutralizing its function.

This multi-label classification will allow us to better characterize drug-target interactions beyond traditional binding affinity prediction and help in:

* **Drug repurposing** – Identifying new uses for existing drugs.
* **Lead optimization** – Selecting the best drug candidates in the early stages.
* **Reducing clinical trial failures** – By improving predictions of drug efficacy and mechanism of action.

By integrating molecular fingerprints, protein sequence embeddings, and interaction graph representations, our model aims to improve the accuracy of DTI predictions, ultimately accelerating the drug discovery process.

* 1. **Methodology**

**Milestone1**

**Step 1: Data Collection & Understanding**

* **Dataset Used:** BindingDB (Kd, IC50, Ki)
* **Kd Dataset:** 52,274 samples with 5 features
* **IC50 Dataset:** 990,630 samples with 5 features
* **Ki Dataset:** 374,820 samples with 5 features
* **Source:** Therapeutic Data Commons (TDC)
* **Attributes:**
  + Drug Information (SMILES strings, molecular descriptors)
  + Drug ID
  + Target Information (Amino acid sequences, protein embeddings)
  + Target ID
  + Experimental Assay Type (Kd, IC50, Ki values)
  + Binding Affinity Values (Y)
  + Interaction Types: (To be obtained from external sources)

**Step 2: Data Preprocessing & Transformation**

* **Insights from Drug\_ID and Target\_ID**: Since the number of **Drug\_IDs** was lower than the number of **Drug Names**, we initially assumed that some Drug\_IDs were missing. However, after extensive research, we discovered that multiple Drug\_IDs could correspond to the same drug, and vice versa, indicating a mapping inconsistency. Ultimately, we determined that **Drug\_IDs lacked meaningful significance**, leading us to drop the column for further analysis. The same approach was applied to **Target ID** as well.

Total unique drugs: 699875  
Total unique targets: 6548

Average targets per drug: 1.77  
Average drugs per target: 188.78

* **Log Transformation:**

A green line graph with numbers

Description automatically generated

* Above image represents the data before applying log transformation, highlighting that the **Y-values are right-skewed** across all three assay types.

A green graph with numbers and numbers

Description automatically generated

* Above image shows the data after applying log transformation, where the values have been **normalized and standardized** across all assay types.
* Applied pY = -log10(Y) transformation to normalize binding affinity values.
* Ensures better model stability and reduces skewness in Kd, IC50, Ki values.
* **Distribution of Drug-Target interactions across assay types:**

A graph of a number of different colored squares

Description automatically generated

* **Distribution of Targets across assay types:**

A graph of a distribution of targets

Description automatically generated

* **Pearson Correlation Heatmap between Assay Types:**

A red and blue squares

Description automatically generated

**Step 3: Data Analysis & Visualization**

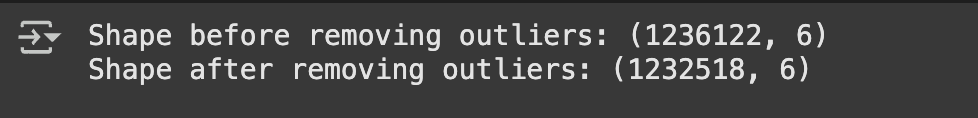
* Boxplots of pY values to check variance across assays.

A graph with a line

Description automatically generated

A green rectangular object with lines

Description automatically generated



* **Initial Exploration (Box Plots & Histograms - Before Transformation)**
  + Identified right-skewness and extreme outliers in Y\_Kd, Y\_IC50, and Y\_Ki.
  + Histograms confirmed data concentration at lower values, indicating potential binning refinements.
* **Data Transformation & Outlier Removal**
  + Applied log transformation to normalize distributions.
  + Removed 3,604 extreme outliers, reducing dataset size from 1,236,122 to 1,232,518 rows while maintaining 6 columns.
* **Post-Processing Visualization (Histograms & Box Plots - After Transformation)**
  + Skewness significantly reduced distributions appeared more normal (especially for Y\_Kd & Y\_IC50).
  + Box plots showed symmetrical distributions with reduced variability.
* **Outcome**
  + The dataset is now cleaned, normalized, and ready for statistical modeling.
  + Ensuring clear documentation of outlier removal criteria is essential for analysis justification.
* **Scatterplots for Drug-Target Interaction Distributions**.

A graph of a drug

Description automatically generated

* The scatter plot is plotted on a log-log scale due to the wide range of interaction counts.
* Curved Trend in Distribution: The distribution appears to follow a power-law behavior, suggesting that a small number of drugs interact with many targets, while most drugs interact with very few targets.
* Polypharmacology Insights: Some drugs are highly polypharmacological, meaning they interact with many targets. These drugs occupy the upper-right portion of the plot, showing high drug interaction counts and high target interaction counts.  
  Other drugs appear highly selective, interacting with very few targets.

Milestone2

Step 4: Feature Engineering

We extract three types of features (Drug, Target, Drug-Target Pair Features) to enhance classification performance.

**4.1 Drug Features**

Chemical Descriptors (RDKit):

* Molecular weight, LogP, hydrogen bond donors/acceptors, rotatable bonds, etc.  
  Fingerprint Representations (ECFP4/Morgan fingerprints):
* Used to capture the structural patterns of molecules.  
  Pre-trained Molecular Embeddings (ChemBERTa, Mol2Vec):
* Transformer-based embeddings for molecules.

**4.2 Target Features**

Physicochemical Properties (Protein-level descriptors):

* Molecular weight, isoelectric point, secondary structure.  
  Pre-trained Protein Sequence Embeddings (ProtTrans, TAPE, UniRep):
* Deep learning embeddings for biological sequences.  
  Protein-Protein Interaction (PPI) Network Features:
* Degree centrality, pathway enrichment for proteins.

**4.3 Drug-Target Pair Features**

* Concatenation of Drug & Target Embeddings
* Interaction Graph Representations (DGL/Deep Graph Networks)

Step 5: Multi-Label Classification Model Development

* Model Types Explored:
  + Traditional ML Models: Random Forest, XGBoost, SVM, Logistic Regression.
  + Deep Learning Models: Multi-Label Neural Networks (MLNN), Transformer-based models.
  + Graph Neural Networks (GNNs): Drug-Target Graph Interaction Prediction using Deep Graph Networks (DGL).
* Why Multi-Label Classification?
  + A drug-target pair may have multiple interaction types (e.g., Activator & Inhibitor).
  + Each drug-target pair can be labeled with multiple interaction types.

Step 6: Model Training & Hyperparameter Tuning

* Split Data into Training, Validation, Test Sets.
* Apply Oversampling/SMOTE for Class Imbalance Handling.
* Optimize Hyperparameters using Grid Search & Bayesian Optimization.

Milestone3

Step 7: Model Evaluation & Interpretation

7.1 Evaluation Metrics:

* F1-Score (per label & overall)
* Precision-Recall (PR) Curve for Multi-Label Classification
* Mean Average Precision (MAP) Score
* Confusion Matrix for Interaction Type Predictions

7.2 Interpretability Techniques:

* SHAP Values (Feature Importance for Prediction).
* Attention Maps (for Transformer Models).

Step 8: Deployment & Real-World Application

* Deploy Model as API using Flask/FastAPI.
* Integration into Drug Discovery Pipelines.
* Visualization Dashboard for Interaction Predictions.

**Expected Outcomes: DTI Multi-Label Classification**

* **Accurate multi-label classification** of drug-target interactions into **Activator, Inhibitor, Partial Agonist, and Antagonist** across diverse assay types (Kd, IC50, Ki).
* **Enhanced feature engineering** using **molecular fingerprints, protein embeddings (ProtTrans, ChemBERTa), and graph-based representations** to improve interaction predictions.
* **Improved assay correlation analysis** to better understand **polypharmacology and target promiscuity** in drug discovery.
* **Robust generalization** of AI models to **unseen drugs & proteins**, aiding in **lead optimization & virtual screening**.
* **Integration with clinical data** to link **DTI predictions with clinical trial success/failure** for **better patient stratification & precision medicine**.
* **Scalability & industrial impact** for **drug repurposing, toxicity prediction, and AI-driven drug discovery pipelines in pharma & biotech**.
* **AI-driven DTI prediction will accelerate drug discovery, reduce failures, and enable next-gen precision medicine.**