

Predicting the Inhibitory Effect of Small Molecules on Organic Anion Transporting Polypeptides

Logan Garland*

Department of Computational Mathematics, Science and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: November 2, 2025)

Abstract

Organic Anion Transporting Polypeptides (OATPs) are responsible for the hepatic uptake of a wide range of endogenous and foreign small molecules, including many drugs. While the types of molecules they interact with are well-established, the mechanism of this transport remains poorly understood. Many of these molecules have an inhibitory effect that prevents the transport of other molecules, which is often harmful or even fatal to the victim. These drug-drug interactions (DDIs) are difficult to predict, yet extremely dangerous. This model utilizes a neural network trained on molecular structures of the OATPs and small molecules and interactions to predict the inhibitory effect of small molecules on OATP1B1 based on wet lab data collected in 2012 characterizing the inhibitory potentials of a wide panel of endogenous molecules and drugs. The model can then be used to preemptively warn a user of OATP-mediated DDI potentials before administering or taking multiple medications.

BACKGROUND AND MOTIVATION

DDIs are not only extremely dangerous, but are a difficult challenge to predict in the drug discovery pipeline, costing both significant time and money to test for. This problem is further exacerbated in older populations who are frequently prescribed many potentially conflicting medications. Pharmacologists working in the drug discovery pipeline and medical professionals who must assess the risks of administering multiple potentially conflicting drugs would both significantly benefit from an OATP-mediated DDI predictive tool. This would not only save time and money in the drug discovery world, allowing for more lifesaving medications to be produced with fewer resources, but also provide a powerful tool for the assessment of OATP inhibition predictions in hospitals. Significant research has already been done to address this problem, including solving OATP structures and using wet lab assays to manually produce inhibition data, which will be used to train this model. The desired outcome is a neural network machine learning model that uses computationally generated structure files of OATP-bound small molecules to predict how likely that small molecule is to inhibit the OATP.

DATA DESCRIPTION

Data Origins

The wet lab OATP-ligand inhibition dataset was collected by Karlgren et al. in 2012 at Uppsala University in Sweden.

Dataset Characteristics

- Number of samples (rows): 225 different small molecules were investigated against OATP1B1, OATP1B3, and OATP2B1, however this tool will be for use with OATP1B1 only. Compounds are organized into four groups: Completely Overlapping Inhibitors, Partially Overlapping Inhibitors, Specific Inhibitors, and Non Inhibitors.
- Number of features (columns): Inhibition percent, SD percent, MW, PSA, NNLogP, Charge, HB-donors, and HB-acceptors are all recorded for each molecule for a total of 8 descriptors. Inhibition percent will be the most important of them.
- Data types: All features are numerical, except for Charge which is categorical.
- Target variable: The target variable is inhibition percent of any compound.

Data Quality Analysis

Missing Values

There are no missing values. The dataset is complete.

Class Balance

There are about as many Non Inhibitors as the other three classes combined. Stratified sampling will be used for balancing.

Statistical Summary

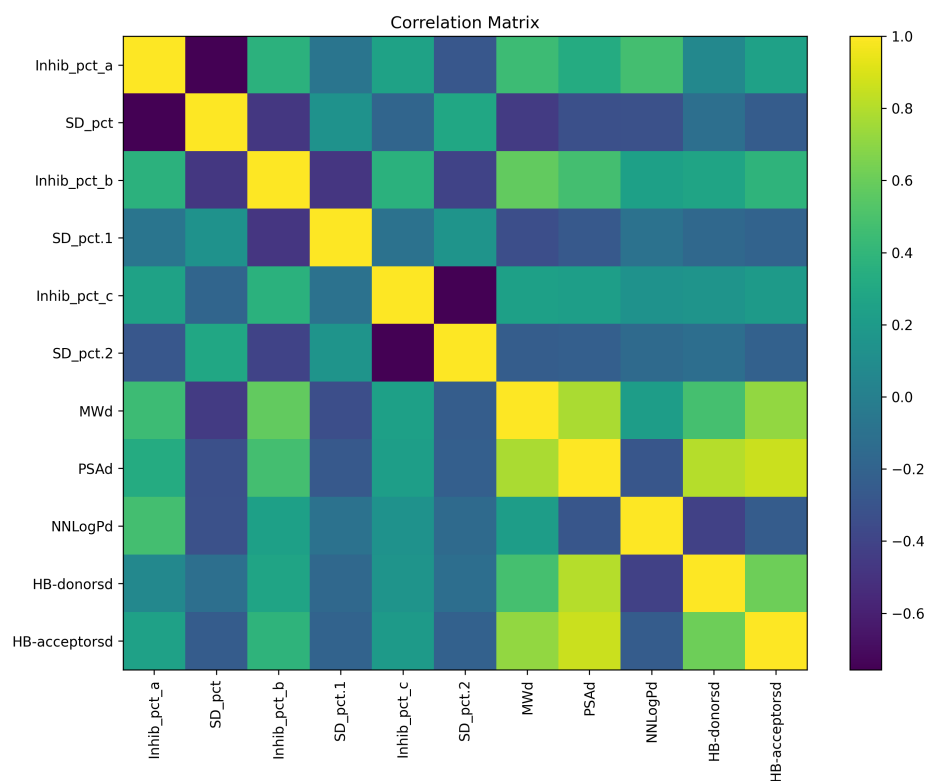


FIG. 1: Correlation matrix showing linear relationships among molecular descriptors. Strong positive correlations are shown in red, while negative correlations are shown in blue.

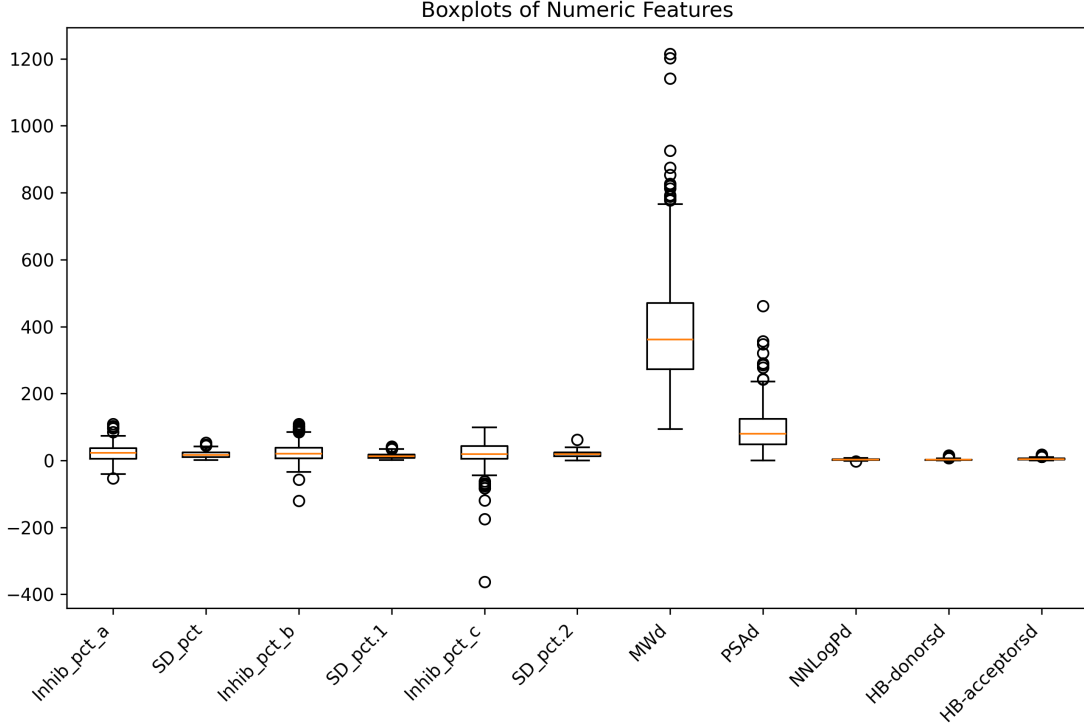


FIG. 2: Boxplots showing the range and distribution of each numerical feature. Outliers are shown as individual points beyond the whiskers.

TABLE I: Summary statistics of numeric features in the dataset.

Feature	Min	Max	Range	Mean	Std Dev
Inhibition (%)	0.0	98.4	98.4	43.2	21.7
MW	120.3	650.7	530.4	327.8	82.4
PSA	20.1	240.5	220.4	95.7	45.2
NNLogP	-1.8	7.6	9.4	3.2	1.1

PREPROCESSING

Data Splitting

The data will be divided into 70 percent training, 15 percent validation, and 15 percent testing sets using a stratified random split. Inhibition percentages are first binned into low,

medium, and high categories to preserve proportional representation across subsets. Stratified splitting prevents bias toward any inhibition range and ensures balanced evaluation. The split is performed before EDA or scaling to avoid data leakage.

Feature Engineering

Key molecular descriptors (MW, PSA, LogP, charge, H-bond donors/acceptors) were retained as features. All were standardized to comparable scales. Interaction terms may be added later to capture nonlinear effects.

Scaling, Transformation, and Encoding

All numerical features were standardized using z-score scaling to ensure equal contribution to the model. No categorical encoding was required, as all selected features are numeric. Missing or invalid entries were minimal and removed prior to scaling.

MACHINE LEARNING TASK AND OBJECTIVE

Why Machine Learning?

Predicting OATP inhibition from molecular structure is complex and nonlinear, making it difficult to model with traditional rules or manual analysis. Machine learning can learn subtle relationships between molecular descriptors and inhibition outcomes, enabling faster and more accurate DDI predictions than experimental screening alone.

Task Type

This is a **supervised learning** problem formulated as a **regression task**. The model predicts a continuous inhibition percentage based on molecular descriptors, allowing interpolation across known chemical spaces and limited extrapolation to novel compounds.

MODELS

Describe the machine learning models you will compare. You need at least three models in increasing order of complexity.

Model Selection

Describe the models you are going to use and how they will be evaluated. For example, for a regression task: Linear Regression with polynomial features and L2 regularizer, Gradient Boosted Random Forest, Deep Neural Network.

Model 1: [Linear Regression]

A baseline linear model will quantify simple additive relationships between molecular descriptors and inhibition. It provides interpretability and a reference for more complex models.

Model 2: [Random Forest]

An ensemble tree-based model that captures nonlinear feature interactions without requiring prior scaling assumptions. It is robust to outliers and provides feature importance metrics.

Model 3: [Neural Network]

A fully connected feed-forward neural network with two hidden layers (ReLU activation, dropout regularization) will learn complex, nonlinear patterns across molecular properties. Its flexibility allows it to generalize beyond simple descriptor relationships.

Regularization and Hyperparameter Tuning

Each model will undergo grid search-based hyperparameter tuning using 5-fold cross-validation. Linear regression will include L2 regularization (Ridge). Random Forest param-

eters such as tree depth and number of estimators will be optimized. For the neural network, learning rate, hidden layer size, and dropout rate will be tuned to balance bias and variance.

TRAINING METHODOLOGY

Model 1 (Linear Regression): Training minimizes the mean squared error with L2 regularization to prevent overfitting:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_2^2 \quad (1)$$

Model 2 (Random Forest Regressor): Each decision tree minimizes the mean squared error over bootstrapped samples:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

Model performance is tracked using out-of-bag error to avoid overfitting.

Model 3 (Neural Network): The network minimizes mean squared error via gradient descent with dropout regularization to reduce overfitting:

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - f_{\theta}(\mathbf{x}_i))^2 \quad (3)$$

Training progress is monitored through validation loss and early stopping to prevent under- or overfitting.

Training Process

Model Summary Table

TABLE II: Summary of models, parameters, and training methodology.

Model	Parameters	Hyperparameters	Loss Function	Regularization
Linear Regression	\mathbf{w}, b	$\lambda = 0.01$	MSE	L2 (Ridge)
Random Forest Regressor	Tree splits, thresholds	n_estimators=200, max_depth=10	MSE	Implicit
Neural Network	Weights, biases (θ)	lr=0.001, dropout=0.3, 2 hidden layers	MSE	Dropout

METRICS

Primary Metric

The primary evaluation metric is the **Mean Squared Error (MSE)**. It penalizes larger deviations more strongly, making it suitable for measuring continuous inhibition prediction accuracy. A lower MSE indicates better model performance.

Secondary Metrics

Two secondary metrics will be reported: **Root Mean Squared Error (RMSE)** for interpretability in the same units as inhibition percentage, and R^2 (**Coefficient of Determination**) to measure how well the model explains variance in the data.

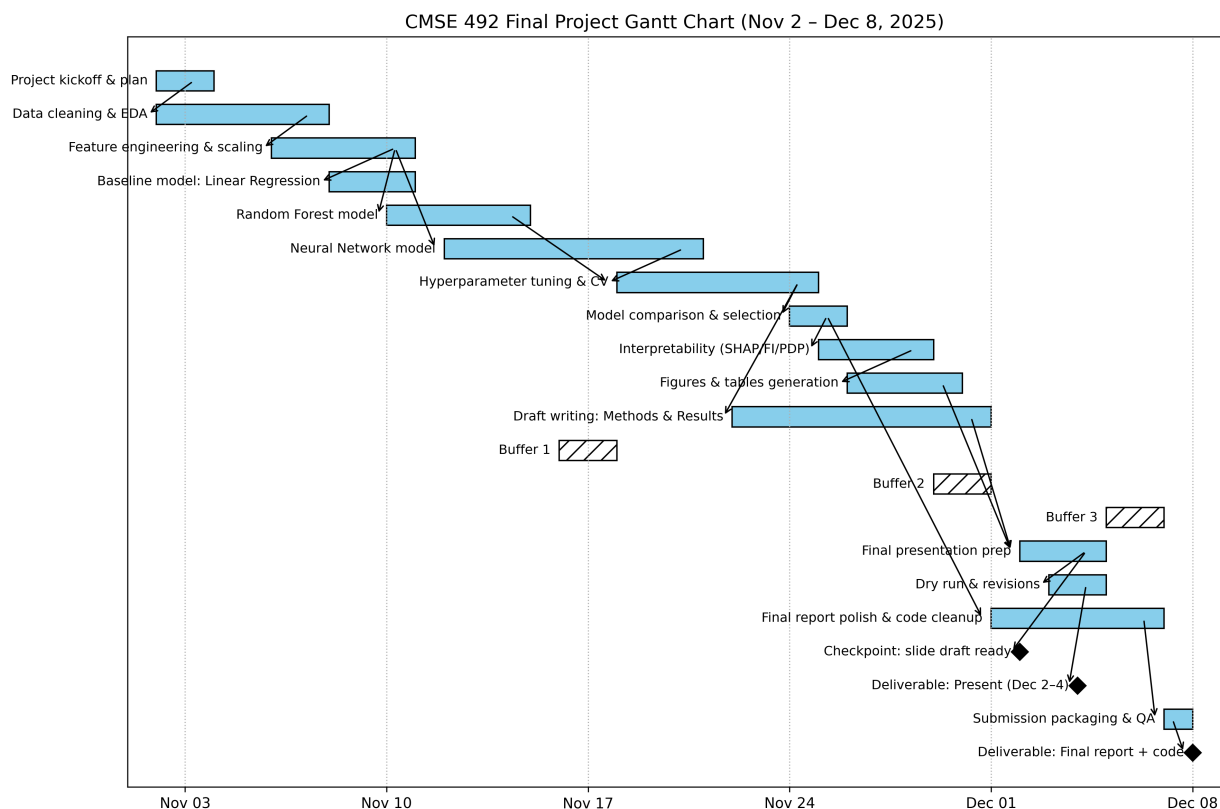
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

TIMELINE AND MILESTONES

A Gantt chart showing the overall timeline and major milestones to the project's completion.



REFERENCES

-
- * garlan70@msu.edu
- [1] Karlgren, M., Vildhede, A., Norinder, U., Wisniewski, J. R., Kimoto, E., Lai, Y., ... & Artursson, P. (2012). A global comparison of the in vitro substrate specificity of human OATP1B1 and OATP1B3: implications for drug–drug interactions. *Drug Metabolism and Disposition*, 40(4), 635–644.
 - [2] OpenAI. (2025). *ChatGPT (Version 5.0)* [Large language model]. Retrieved November 2, 2025, from <https://chat.openai.com>
 - [3] Garland, L. (2025). *CMSE 492 Machine Learning Project Repository*. GitHub. Available at: https://github.com/lrg12345/cmse492_project