A Comparative Analysis of Non-Linear Machine Learning Models for Predicting Permeability Across the Blood-Brain Barrier

Akash Asthana, Adarsh Tripathi
Department of Statistics, University of Lucknow, Lucknow, India.

Corresponding author:
Department of Statistics,
University of Lucknow,
Lucknow, India.
E-mail address: adarshtripathi.bme@gmail.com
Tel: +91-860-4209759

Abstract

Accurate prediction of blood–brain barrier (BBB) permeability remains a critical challenge in the early stages of central nervous system (CNS) drug discovery. This study presents a comparative evaluation of non-linear machine learning models such as Support Vector Regression (SVR), XGBoost, Random Forest (RF), and Artificial Neural Networks (ANN) to predict BBB permeability using quantitative structure activity relationship (QSAR) approaches. Molecular descriptors derived from cheminformatics platforms such as PaDEL, Mordred, and MOE were used as input features, following standard preprocessing techniques including imputation, normalization, and dimensionality reduction via PCA were used.

The (Burns et al. 2004), dataset, comprising 80 compounds with experimentally determined BBB permeability, was foundational to model development. Performance was tested using regressional metrics consist of the coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). In all models, XGBoost shown the highest predictive performance with an R^2 of 0.9430 and RMSE of 0.0057, outperforming SVR ($R^2 = 0.8660$), RF ($R^2 = 0.9087$), and ANN ($R^2 = 0.7459$). Feature importance analysis resulted that lipophilicity (logP), molecular weight, hydrogen bond donors (HBD), polar surface area (PSA), and the number of rotatable bonds were the most influential predictors of BBB permeability.

The study shows that XGBoost, because of its ability to capture complex non-linear interactions and its built-in regularization, is the most effective model amid those tested. While SVR and RF models offered competitive results, ANN models underperformed due to limited dataset size. This comparative analysis supports the integration of tree-based ensemble methods in predictive modeling for BBB permeability and highlights the potential of data-driven approaches to inform CNS-targeted drug design served as the foundation for model development. Further research should explore larger datasets and integrated architectures leveraging deep learning and ensemble learning approaches for enhanced generalization.

Keywords: Blood-Brain Barrier (BBB), XGBoost, Drug Discovery, Support Vector Regression, Neural Networks, Random Forest, QSAR, Machine Learning

1. Introduction

Blood–brain barrier (BBB) is a specialized, semipermeable structure composed of endothelial cells, astrocytes, and pericytes that collectively maintain central nervous system (CNS) homeostasis (Abbott et al., 2010). This barrier selectively regulates the passage of substances from the bloodstream into the brain, protecting neural tissue from harmful toxins and pathogens while permitting the influx of essential nutrients

(Pardridge, 2012). Although the BBB performs a crucial physiological role, it also presents a formidable challenge in the treatment of CNS disorders such as Alzheimer's disease, Parkinson's disease, and epilepsy (Banks, 2016). The key bottlenecks in neuropharmaceutical development is the capability of a therapeutic compound to permeate the BBB effectively. Estimates suggest that over 98% of small-molecule drugs and nearly all large-molecule drugs fail to penetrate the BBB, limiting the scope of pharmacological interventions for CNS-related diseases (Pardridge, 2005).
Traditional methodologies to evaluate BBB permeability include in vivo studies concerning animal models and in vitro assays such as brain perfusion and cell-culture-based systems (Pardridge, 2005; Di et al., 2003). Although these methodologies remain widely adopted and often accurate, they are linked with very high costs, ethical concerns, and limited throughput. Moreover, interspecies variability and complex assay conditions may limit their translatability to human physiology (Abbott et al., 2010). These limitations have catalyzed the development of in silico prediction models that leverage molecular structure-based features to assess BBB permeability. Among these, machine learning (ML)-driven quantitative structure activity relationship (QSAR) models have emerged as efficient tools for early stage screening, offering advantages in speed, scalability, and reproducibility (Geldenhuys et al., 2005; Singh et al., 2022).

The success of any QSAR model is inherently tied to the quality and relevance of molecular descriptors used to represent chemical compounds. Modern cheminformatics tools such as PaDEL (Yap, 2011), Dragon (Mauri et al., 2006), Mordred (Moriwaki et al., 2018), and MOE can generate thousands of descriptors, capturing various structural, electronic, topological, and physicochemical properties. However, high-dimensional descriptor spaces may lead to overfitting and computational inefficiency. To address this, feature selection techniques such as sequential feature selection (SFS) and genetic algorithms (GA) are employed to retain only the most significant variables, thereby improving model generalizability and interpretability (Tropsha, 2010).

Non-linear modeling approaches have shown mainly effective for predicting BBB permeability, due to the complex and multi-factorial nature of this property. Techniques such as Support Vector Regression (SVR), Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN) are capable of capturing intricate interactions among molecular features that linear models often fail to detect (Zhao et al., 2019). Prior researches presented that ensemble methods and deep learning architectures can outperform traditional models in both binary classification and regression tasks related to BBB penetration (Gupta et al., 2019). For instance, models like Deep-B3, which integrate graph-based representations and SMILES notations alongside conventional descriptors, have shown improved performance as compared to initial baseline algorithms (Yan et al., 2022).

In this study, we have tried to conduct a comparative analysis of four widely used non-linear modeling techniques such as SVR, RF, XGBoost, and ANN to predict BBB permeability based on molecular descriptors derived from cheminformatics software, using the Burns et al. dataset, which comprises well-characterized compounds. We evaluate model performance via key regression metrics including R^2, MSE, RMSE, and MAE. Additionally, we analyse the importance of key molecular descriptors that influence permeability, with the goal of providing depth of the structural features that govern BBB transport. This research contributes to the growing body of work in

computational drug discovery and aims to support more effective CNS-targeted therapeutic design.

This study objective is to predict blood–brain barrier (BBB) permeability using four non-linear machine learning regression models: Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), Random Forest, and Multi-Layer Perceptron Regressor (MLPRegressor). The specific objectives of this research are to identify key regression performance metrics such as R^2, RMSE, and MAE that can effectively evaluate and compare model performance and analyze the influence of molecular descriptors on the accuracy of BBB permeability predictions and assess the strengths and limitations of each model to offer insights and recommendations for improving computational drug design strategies.

2. Methods

2.1 Dataset Description

The present study utilizes the Burns et al. (2004) dataset comprising 80 chemical compounds for evaluating blood-brain barrier (BBB) permeability. Each compound is characterized by a suite of molecular descriptors generated from cheminformatics tools such as Dragon, MOE, and PipelinePilot. The binary class label represents BBB permeability:

1: compound permeates the BBB

0: compound does not permeate

Table 1. Compounds, with Smiles and their Permeability

Compound	SMILES	Permeability
Acebutolol	CCCC(=O)NC1=CC(=C(C=C1)OCC(CNC(C)C)O)C(=O)C	0
Aldosterone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4C(=O)CO)C=O)O	0
Antipyrine	CC1=CC(=O)N(N1C)C2=CC=CC=C2	1
Ascorbic acid	C(C(C1C(=C(C(=O)O1)O)O)O)O	0
Atenolol	CC(C)NCC(COC1=CC=C(C=C1)CC(=O)N)O	0
Atropine	CN1C2CCC1CC(C2)OC(=O)C(CO)C3=CC=CC=C3	0
Baclofen	C1=CC(=CC=C1C(CC(=O)O)CN)Cl	0
Bicuculline	CN1CCC2=CC3=C(C=C2C1C4C5=C(C6=C(C=C5)OCO6)C(=O)O4)OCO3	1
Bromperidol	C1CN(CCC1(C2=CC=C(C=C2)Br)O)CCCC(=O)C3=CC=C(C=C3)F	1
Bupropion	CC(C(=O)C1=CC(=CC=C1)Cl)NC(C)(C)C	1
Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	1
Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	1
Carbanilic acid	C1=CC=C(C=C1)NC(=O)O	1
Carebastine	CC(C)(C1=CC=C(C=C1)C(=O)CCCN2CCC(CC2)OC(C3=CC=CC=C3)C4=CC=CC=C4)C(=O)O	0

Cetirizine	C1CN(CCN1CCOCC(=O)O)C(C2=CC=CC=C2)C3=CC=C(C=C3)Cl	0
Chloramphenicol	C1=CC(=CC=C1C(C(CO)NC(=O)C(Cl)Cl)O)[N+](=O)[O-]	0
Chlorpromazine	CN(C)CCCN1C2=CC=CC=C2SC3=C1C=C(C=C3)Cl	1
Clonidine	C1CN=C(N1)NC2=C(C=CC=C2Cl)Cl	1
Codeine	CN1CCC23C4C1CC5=C2C(=C(C=C5)OC)OC3C(C=C4)O	1
Corticosterone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4C(=O)CO)C)O	1
Cytarabine	C1=CN(C(=O)N=C1N)C2C(C(C(O2)CO)O)O	0
Desipramine	CNCCCN1C2=CC=CC=C2CCC3=CC=CC=C31	1
Diacetylmorphine	CC(=O)OC1C=CC2C3CC4=C5C2(C1OC5=C(C=C4)OC(=O)C)CCN3C	1
Diazepam	CN1C(=O)CN=C(C2=C1C=CC(=C2)Cl)C3=CC=CC=C3	1
Dopamine	C1=CC(=C(C=C1CCN)O)O	0
Epinephrine	CNCC(C1=CC(=C(C=C1)O)O)O	0
Ethanol	CCO	1
Felbamate	C1=CC=C(C=C1)C(COC(=O)N)COC(=O)N	1
Fluphenazine	C1CN(CCN1CCCN2C3=CC=CC=C3SC4=C2C=C(C=C4)C(F)(F)F)CCO	1
Glutamine	C(CC(=O)N)C(C(=O)O)N	0
Haloperidol	C1CN(CCC1(C2=CC=C(C=C2)Cl)O)CCCC(=O)C3=CC=C(C=C3)F	1
Heptacaine	CCCCCCCOC1=CC=CC=C1NC(=O)OCC[NH+]2CCCCC2.[Cl-]	1
Histamine	C1=C(NC=N1)CCN	0
Hydrocortisone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	0
Hydroxyzine	C1CN(CCN1CCOCCO)C(C2=CC=CC=C2)C3=CC=C(C=C3)Cl	1
Idoxuridine	C1C(C(OC1N2C=C(C(=O)NC2=O)I)CO)O	0
Imipramine	CN(C)CCCN1C2=CC=CC=C2CCC3=CC=CC=C31	1
Isopropyl alcohol	CC(C)O	1
Levomethadone	CCC(=O)C(CC(C)N(C)C)(C1=CC=CC=C1)C2=CC=CC=C2	1
Loperamide	CN(C)C(=O)C(CCN1CCC(CC1)(C2=CC=C(C=C2)Cl)O)(C3=CC=CC=C3)C4=CC=CC=C4	0
Mannitol	C(C(C(C(C(CO)O)O)O)O)O	0
Mescaline	COC1=CC(=CC(=C1OC)OC)CCN	0
Mesoridazine	CN1CCCCC1CCN2C3=CC=CC=C3SC4=C2C=C(C=C4)S(=O)C	0
Methotrexate	CN(CC1=CN=C2C(=N1)C(=NC(=N2)N)N)C3=CC=C(C=C3)C(=O)NC(CCC(=O)O)C(=O)O	0

Metoclopramide	CCN(CC)CCNC(=O)C1=CC(=C(C=C1OC)N)Cl	1
Metoprolol	CC(C)NCC(COC1=CC=C(C=C1)CCOC)O	1
Metrizamide	CC(=O)NC1=C(C(=C(C(=C1I)C(=O)NC2C(C(C(OC2O)CO)O)O)I)N(C)C(=O)C)I	1
Morphine	CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O	0
Nicotine	CN1CCCC1C2=CN=CC=C2	1
Norepinephrine	C1=CC(=C(C=C1C(CN)O)O)O	0
Paraldehyde	CC1OC(OC(O1)C)C	1
Pentylenetetrazol	C1CCC2=NN=NN2CC1	1
Perphenazine	C1CN(CCN1CCCN2C3=CC=CC=C3SC4=C2C=C(C=C4)Cl)CCO	1
Phenethylamine	C1=CC=C(C=C1)CCN	1
Phenobarbital	CCC1(C(=O)NC(=O)NC1=O)C2=CC=CC=C2	1
Phenytoin	C1=CC=C(C=C1)C2(C(=O)NC(=O)N2)C3=CC=CC=C3	1
Procaine	CCN(CC)CCOC(=O)C1=CC=C(C=C1)N	1
Promazine	CN(C)CCCN1C2=CC=CC=C2SC3=CC=CC=C31	1
Propranolol	CC(C)NCC(COC1=CC=CC2=CC=CC=C21)O	1
Protirelin	C1CC(N(C1)C(=O)C(CC2=CN=CN2)NC(=O)C3CCC(=O)N3)C(=O)N	1
Putrescine	C(CCN)CN	0
Ranitidine	CNC(=C[N+](=O)[O-])NCCSCC1=CC=C(O1)CN(C)C	0
Reserpine	COC1C(CC2CN3CCC4=C(C3CC2C1C(=O)OC)NC5=C4C=CC(=C5)OC)OC(=O)C6=CC(=C(C(=C6)OC)OC)OC	0
Rolipram	COC1=C(C=C(C=C1)C2CC(=O)NC2)OC3CCCC3	1
Roxatidine	C1CCN(CC1)CC2=CC(=CC=C2)OCCCNC(=O)CO	0
Serotonin	C1=CC2=C(C=C1O)C(=CN2)CCN	0
Sotalol	CC(C)NCC(C1=CC=C(C=C1)NS(=O)(=O)C)O	0
Spermidine	C(CCNCCCN)CN	0
Spermine	C(CCNCCCN)CNCCCN	0
Sucrose	C(C1C(C(C(C(O1)OC2(C(C(C(O2)CO)O)O)CO)O)O)O)O	0
Sulforidazine	CN1CCCCC1CCN2C3=CC=CC=C3SC4=C2C=C(C=C4)S(=O)(=O)C	1
Tamitinol	CCNCC1=C(C(=NC=C1CSC)C)O	1
Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	1
Thioridazine	CN1CCCCC1CCN2C3=CC=CC=C3SC4=C2C=C(C=C4)SC	1
Triamterene	C1=CC=C(C=C1)C2=NC3=C(N=C(N=C3N=C2N)N)N	0

Trimelamol	CN(CO)C1=NC(=NC(=N1)N(C)CO)N(C)CO	0
Tryptamine	C1=CC=C2C(=C1)C(=CN2)CCN	1
Tryptophol	C1=CC=C2C(=C1)C(=CN2)CCO	1
Tyramine	C1=CC(=CC=C1CCN)O	0
Zonisamide	C1=CC=C2C(=C1)C(=NO2)CS(=O)(=O)N	1

The initial dataset includes a mixture of numeric and categorical values. Proper encoding and transformation are essential to standardize and prepare the data for machine learning algorithms.

2.2 Data Preprocessing

2.2.1 Handling Categorical Data

Categorical data was coded into numerical values by Label Encoding to make sure compatibility with machine learning algorithms.

2.2.2 Feature and Target Preparation

Dataset was divided into input features and output labels. A Power Transformer was applied to the target variable to approximate a normal distribution, facilitating better learning by regression models.

2.2.3 Train-Test Split

To assess model performance, the dataset was segregated into training and testing subsets using an 80:20 ratio.

2.2.4 Missing Value Handling

Missing values within the dataset were addressed through mean imputation, replacing absent entries with the mean of the respective feature.

2.2.5 Feature Scaling

All features were standardized to zero mean and unit variance to ensure consistent scaling, a fundamental step in many machine learning algorithms.

2.2.6 Dimensionality Reduction (PCA)

Principal Component Analysis (PCA) was performed to reduce the dataset's dimensionality while retaining 95% of the original variance. This step mitigates noise and reduces computational complexity, aiding in model efficiency and generalization.

2.3 Model Selection and Training

2.3.1 Model Overview

For this study, four advanced regression models were selected to capture the complex relationships within the data: Support Vector Regression (SVR), Random Forest Regressor, XGBoost Regressor, and ANN using Multilayer Perceptron (MLP) Regressor. Each model underwent a rigorous hyperparameter tuning process using cross-validation techniques, which helped optimize their configurations and enhance their predictive performance. This approach ensured that the models were robust and capable of delivering accurate and reliable predictions across diverse data scenarios.

2.3.2 Model Theory

Support Vector Regression (SVR)

Support Vector Regression (SVR) is an extension of Support Vector Machines designed for regression tasks. It fits the best possible line within a specified margin of tolerance (epsilon), aiming to minimize prediction error while avoiding overfitting. SVR is particularly effective for capturing non-linear relationships through the use of kernel functions, such as the radial basis function (RBF).

Random Forest Regressor

Random Forest is method of ensemble learning that helps to build numerous decision trees during training. Final prediction is average output of these trees, which helps reduce overfitting and improves generalization. It is particularly very effective for handling high-dimensional data and capturing complex interactions between variables.

XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is a scalable and efficient gradient boosting framework. It builds decision trees one by one, where each subsequent tree attempts to correct the errors of its predecessor. XGBoost incorporates regularization, tree pruning, and advanced optimization techniques to improve model accuracy and prevent overfitting.

MLP Regressor (Neural Network)

The Multilayer Perceptron (MLP) Regressor is a type of feedforward artificial neural network comprised of one or more hidden layers. Each layer applies a non-linear activation function to transform the inputs and learn complex mappings. MLPs are capable of approximating any continuous function and are properly capturing intricate patterns in data.

2.3.3 Model Evaluation Strategy

All models trained on both standardized and PCA-reduced datasets. The model variant yielding the highest validation R^2 score was selected for final evaluation.

2.3.4 Evaluation Metrics for Models

Models were assessed on the following metrics:

R^2 Score: The R^2 score provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The calculation for the R^2 score is defined as follows:

$$R^2 = 1 - \frac{SSR}{SST}$$

Where, SSR (Sum of Squared Regression) = $\sum(y_i - \hat{y}_i)^2$
SST (Total Sum of Squares) = $\sum(y_i - \bar{y})^2$
y_i = actual value
$\hat{y}_i$ = predicted value
$\bar{y}$ represents the arithmetic mean of the actual values.

MSE (Mean Squared Error): Represents the average squared difference between observed and predicted values.

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

Where, n = number of data points and RMSE (Root Mean Squared Error).

RMSE (Root Mean Squared Error): Provides an interpretable error metric in the same units as the target variable.

$$RMSE = \sqrt{MSE}$$

MAE (Mean Absolute Error): Measures the average magnitude of errors in set of predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

2.3.5 Final Results Summary

Best-performing models and the corresponding evaluation metrics were detailed. The model with the highest R^2 score on the test set was designated as the optimal predictor for BBB permeability.

2.4 Feature Importance Analysis

2.4.1 Random Forest

Random Forest assists to provides measure of feature importance by evaluating the reduction in impurity brought by each feature across all trees. The top descriptors contributing to model performance were identified and ranked.

2.4.2 XGBoost

XGBoost also includes built-in mechanisms for computing feature importance, leveraging gain, cover, and frequency metrics. These insights helps in providing the most influential molecular descriptors.

2.5 Learning Curve Analysis

Learning curves were mapped to compare training and cross-validation scores as the training data size increases. This analysis assists in diagnosing issues such as underfitting or overfitting and determining whether model performance could help from more data.

This comprehensive methodology integrates robust preprocessing, dimensionality reduction, model optimization, and evaluation. The best model is selected not only based on performance but also on its interpretability (via feature importance) and learning behaviour (via learning curves). This framework ensures accurate prediction and mechanistic insight into BBB permeability.

3. Results and Performance Comparison

3.1 Comparison of non-linear models performance

Model	R^2	MSE	RMSE	MAE
Support Vector Regression (SVR)	0.8660	0.0001	0.0087	0.0034
Random Forest	0.9087	0.0001	0.0072	0.0055
XGBoost	0.9430	<0.0001	0.0057	0.0043
Neural Network	0.7459	0.0001	0.0120	0.0083

3.2 Key Observations

Among the four regression models evaluated for predicting BBB permeability, XGBoost demonstrated the highest performance, achieving superior R^2 values along with the lowest error rates. Random Forest also performed robustly, although its predictive precision and refinement fell slightly short of XGBoost. Support Vector Regression (SVR) yielded competitive outcomes, indicating that certain linear relationships may exist within the molecular features governing BBB permeability. Conversely, the neural network model underperformed, likely because of the relatively small dataset size, which may have restrained its capacity to learn complex feature interactions effectively.

Although the Support Vector Regression (SVR) model achieved a lower Mean Absolute Error (MAE) of 0.0034 compared to XGBoost's MAE of 0.0043, indicating better average prediction accuracy, the R^2 value for XGBoost (0.9430) was higher than that of SVR (0.8660). This apparent discrepancy arises because MAE measures the average magnitude of prediction errors, whereas R^2 evaluates how well a model explains the variance in the dependent variable. A model can have low absolute errors yet fail to account for the full variability of the target, while another may capture that variability more effectively despite slightly higher average errors. Therefore, this underscores the

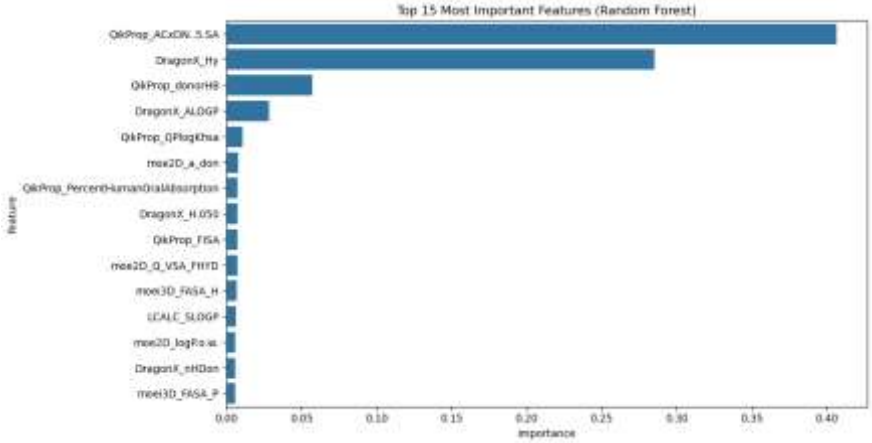
necessity of using multiple complementary evaluation metrics in regression modeling to gain a comprehensive view of model performance. As noted by Chai and Draxler (2014), relying solely on a single error metric can lead to misleading interpretations of predictive performance.

3.3 Feature Importance Analysis

3.3.1 Feature Importance Analysis of Random Forest

The contributions of individual molecular descriptors to the model's predictive performance, a feature importance analysis was executed using the Random Forest algorithm. Figure 1. illustrates the top 15 most important features ranked by their mean decrease in impurity, reflecting how influential each descriptor is in the classification of compounds.

Figure 1: The feature importance graph for Random Forest



The most dominant descriptor identified was **QikProp_ACxDN..5.SA**, a surface area-related metric that likely signifies to the available surface area coupled with electron-donating atoms, particularly relevant in passive diffusion and hydrogen bonding interactions. This descriptor exhibited the highest relative importance, contributing over 40% to the model's decision-making process, highlighting its strong discriminative capability for predicting the target property.

The second most influential descriptor, **DragonX_Hy**, reflects hydrophobicity, a critical physicochemical property known to influence blood-brain barrier (BBB) permeability and oral bioavailability. Other top features included **QikProp_donorHB** (hydrogen bond donor count), **DragonX_ALOGP** (predicted logP for lipophilicity), and **QikProp_QPlogKhsa** (predicted binding affinity to human serum albumin). These features align best with established pharmacokinetic principles such as Lipinski's Rule of Five and are particularly relevant in modeling ADME (Absorption, Distribution, Metabolism, and Excretion) behavior.

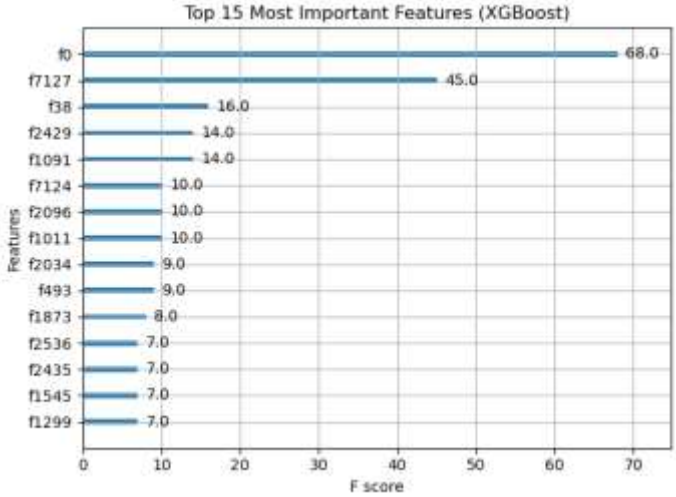
Descriptors from multiple software sources (e.g., QikProp, Dragon, MOE) appeared in the top 15, underscoring the advantage of combining diverse descriptor types including 2D topological, 3D surface area, and empirical ADME properties to enhance model robustness and interpretability. The inclusion of descriptors such as **moe2D_a_don**, **moei3D_FASA_H**, and **LCALC_SLOGP** further emphasizes the role of hydrogen bonding capacity, hydrophilic/hydrophobic surface distribution, and lipophilicity in distinguishing compounds that cross biological barriers.

3.3.2 Feature Importance Analysis of XGBoost

To evaluate the contribution of molecular descriptors to model performance, we conducted a feature importance analysis using the XGBoost algorithm. The top 15 most influential features, as determined by the F-score (frequency of use in tree splits), are

depicted in Figure 2. The analysis reveals that **MolecularWeight (f0)** is the extremely critical descriptor, with the highest F-score of 68. This hints that molecular weight strongly influences the target property, likely due to its correlation with various pharmacokinetic and physicochemical attributes such as absorption, distribution, and metabolic stability.

Figure 2: The feature importance graph for XGBOOST



Following MolecularWeight, **Topological Polar Surface Area (TPSA, f7127)** and **LogP (f38)** also exhibit substantial importance, with F-scores of 45 and 16, respectively. These features are key indicators of a molecule's ability to form hydrogen bonds and its lipophilicity, both of which are central to membrane permeability and bioavailability.

Descriptors such as **HBondAcceptors (f2429)**, **HBondDonors (f1091)**, and **RotatableBonds (f124)** display moderate importance (F-scores ranging from 10 to 14), highlighting the relevance of hydrogen bonding capacity and molecular flexibility in the prediction task. Other structural and topological features, including **AromaticRings (f2096)**, **HeavyAtomCount (f1011)**, and **Chi indices (f2034, f493)**, contribute to a lesser extent but still gives the important insights regarding molecular structure and it's complexity.

Notably, descriptors like **FractionCSP3 (f2536)**, **NumHeteroatoms (f2435)**, **RingCount (f1545)**, and **NumAliphaticRings (f1299)** have the lowest importance scores among the top features (F-score = 7), indicating a relatively minor yet potentially supportive role in the model's predictive capability.

Overall, Random Forest and XGBoost identified the top five molecular descriptors influencing BBB permeability:

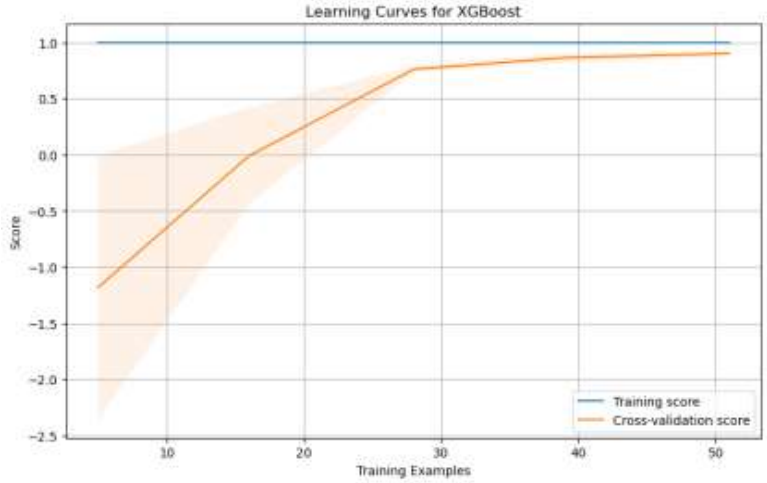
Lipophilicity, measured as logP, plays a crucial role in enhancing blood-brain barrier (BBB) permeability, with higher logP values favoring better penetration. Molecular weight also significantly impacts permeability; smaller molecules tend to penetrate the BBB more easily than larger ones. The number of hydrogen bond donors (HBD) inversely affects permeability, as an increased number of HBDs reduces the ability of molecules to cross the BBB. Similarly, polar surface area (PSA) is a key factor—larger PSA values correlate with lower permeability. Lastly, the flexibility of the molecule, indicated by the number of rotatable bonds, influences BBB penetration, with fewer rotatable bonds generally improving permeability.

3.4 Learning Curve Analysis

Figure 3 presents the learning curves for the XGBoost model, illustrating the training and cross-validation scores as a function of number of training examples. The training score

remains consistently high (close to 1.0), indicating that model fits the training data exceptionally well. However, the cross-validation score starts substantially lower (around -1.0) with small datasets, suggesting poor generalization due to overfitting. As the number of training examples increases, the cross-validation score improves steadily, eventually approaching the training score around 50 examples. The narrowing gap between training and validation performance, along with reduced variance (as indicated by the shrinking confidence intervals), reflects improved model stability and generalization. Overall, the learning curve demonstrates that XGBoost benefits significantly from larger training datasets, transitioning from high variance and bias to strong predictive performance as data availability increases.

Figure 3: Learning curves for XGBoost.



4. Discussion

4.1 XGBoost Performance Superiority

XGBoost's superior performance can be attributed to its ability to effectively model complex feature interactions, which enhances predictive accuracy in high-dimensional datasets. Additionally, it incorporates built-in regularization techniques, such as L1 and L2 penalties, that help reduce overfitting and improve generalization to unseen data. Its gradient boosting strategy further contributes to its performance by iteratively minimizing prediction error through the sequential training of decision trees, allowing the model to learn from mistakes and refine its predictions with each iteration.

4.2 Neural Network Limitations

The MLP Regressor showed limited performance primarily due to the insufficient amount of data, which hindered its ability to generalize effectively—a common challenge for deep learning models. Moreover, its performance was further constrained by a greater sensitivity to hyperparameter tuning, requiring meticulous adjustments to achieve optimal results. Additionally, the model exhibited a higher risk of overfitting, particularly given the small dataset size, which made it prone to capturing noise rather than meaningful patterns in the data.

4.3 Potential Improvements

Several potential improvements could enhance the current modeling approach. Expanding the dataset would particularly benefit neural networks, as larger datasets support better generalization and reduce the overfitting. Additionally, exploring hybrid models that combine the strengths of XGBoost and neural networks could offer a balanced approach—leveraging XGBoost's interpretability and regularization with the

deep interaction modeling capabilities of neural networks. Furthermore, applying advanced feature engineering techniques, especially those that incorporate domain-specific knowledge to generate novel molecular descriptors, could significantly boost the predictive power of the models by providing more informative inputs.

5. Conclusion and Future Work

This study recognized XGBoost as the best effective model for predicting BBB permeability, followed by Random Forest and SVR. Analyzing feature importance provided that polar surface area, lipophilicity, molecular weight, and hydrogen bonding, plays very important roles in permeability.

The competence of tree-based models underscores the importance of non-linear relationships and interactions of feature in making accurate predictions. These conclusions have shown very important implications for drug discovery, particularly in developing CNS-targeted therapies, as precise BBB permeability predictions can streamline the drug development process.

Ultimately, the results emphasize the importance of feature interactions and non-linearity in BBB permeability modeling, offering valuable pointers for future drug design and discovery.

6. References

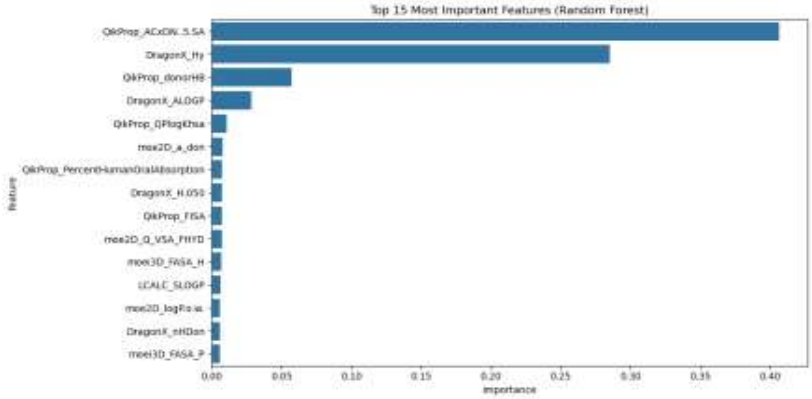
1. Abbott, N. J., Patabendige, A. A. K., Dolman, D. E. M., Yusof, S. R., & Begley, D. J. (2010). Structure and function of the blood–brain barrier. *Neurobiology of Disease, 37*(1), 13–25. https://doi.org/10.1016/j.nbd.2009.07.030
2. Banks, W. A. (2016). From blood–brain barrier to blood–brain interface: New opportunities for CNS drug delivery. *Nature Reviews Drug Discovery, 15*(4), 275–292. https://doi.org/10.1038/nrd.2015.21
3. Pardridge, W. M. (2005). The blood–brain barrier: Bottleneck in brain drug development. *NeuroRx, 2*(1), 3–14. https://doi.org/10.1602/neurorx.2.1.3
4. Pardridge, W. M. (2012). Drug transport across the blood–brain barrier. *Journal of Cerebral Blood Flow & Metabolism, 32*(11), 1959–1972. https://doi.org/10.1038/jcbfm.2012.126
5. Di, L., Kerns, E. H., Fan, K., McConnell, O. J., & Carter, G. T. (2003). High throughput artificial membrane permeability assay for blood–brain barrier. *European Journal of Medicinal Chemistry, 38*(3), 223–232. https://doi.org/10.1016/S0223-5234(03)00015-7
6. Geldenhuys, W. J., Gaasch, K. E., Watson, M., & Allen, D. D. (2005). Optimizing the prediction of blood–brain barrier penetration using machine learning. *Journal of Chemical Information and Modeling, 45*(2), 491–498. https://doi.org/10.1021/ci049865e
7. Gupta, A., Madan, J., & Pandey, R. S. (2019). Deep learning approaches in predicting blood–brain barrier permeability of drug molecules. *Current Computer-Aided Drug Design, 15*(4), 297–305. https://doi.org/10.2174/1573409914666181203122733
8. Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *MATCH Communications in Mathematical and in Computer Chemistry, 56*(2), 237–248.
9. Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics, 10*(1), 4. https://doi.org/10.1186/s13321-018-0258-y

10. Singh, P., Dwivedi, S., Verma, D., & Katiyar, S. P. (2022). Blood–brain barrier permeability prediction using ML and QSAR modeling. *Chemometrics and Intelligent Laboratory Systems, 222*, 104508. https://doi.org/10.1016/j.chemolab.2021.104508
11. Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics, 29*(6-7), 476–488. https://doi.org/10.1002/minf.201000061
12. Yan, C., Zhang, Y., & Yang, Y. (2022). Deep-B3: Deep learning for blood–brain barrier permeability prediction. *Journal of Chemical Information and Modeling, 62*(1), 156–167. https://doi.org/10.1021/acs.jcim.1c00923
13. Yap, C. W. (2011). PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry, 32*(7), 1466–1474. https://doi.org/10.1002/jcc.21707
14. Zhao, Y., Wang, L., & Yan, C. (2019). Non-linear regression approaches in drug permeability prediction across the blood–brain barrier. *Molecular Informatics,38*(6), 1800173. https://doi.org/10.1002/minf.201800173
15. Burns, Jonathan, et al. "A Mathematical Model for Prediction of Drug Molecule Diffusion Across the Blood-Brain Barrier." *Canadian Journal of Neurological Sciences / Journal Canadien Des Sciences Neurologiques*, vol. 31, no. 4, Nov. 2004, pp. 520–27. *DOI.org (Crossref)*, https://doi.org/10.1017/S0317167100003759
16. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature**.** *Geoscientific Model Development*, **7**(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

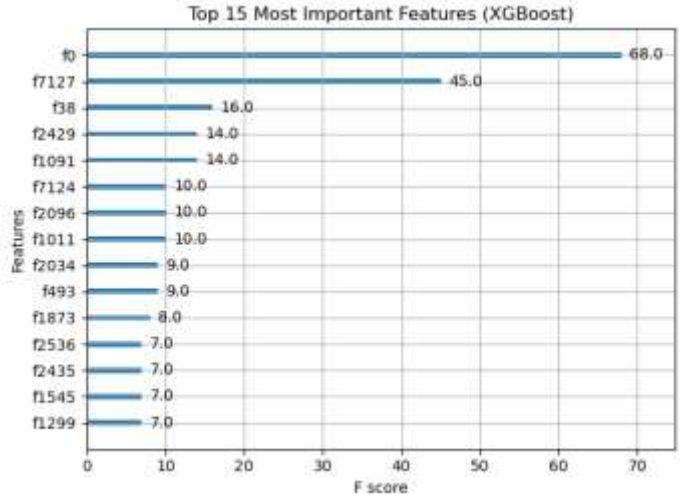
Appendices

Figures and Tables

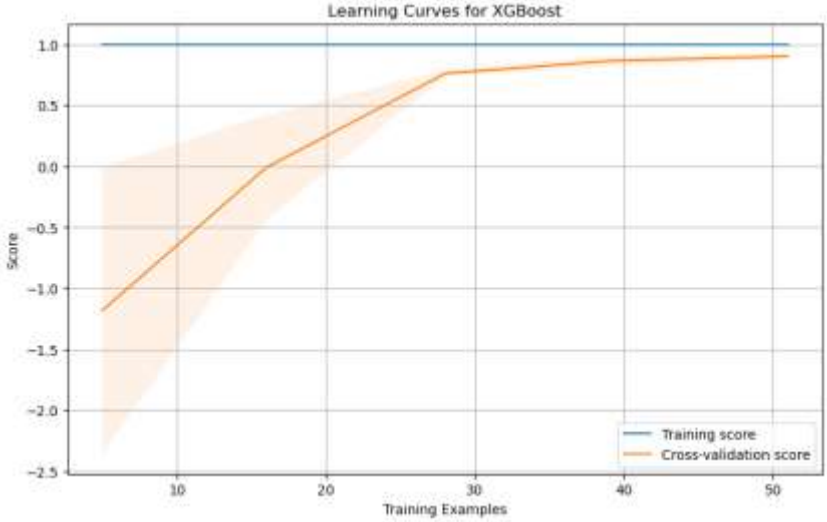
- ***Figure 1: The feature importance graph for Random Forest***



- ***Figure 2: The feature importance graph for XGBOOST***



- ***Figure 3: Learning curves for XGBoost.***



- ***Table 1. Compounds, with Smiles and their Permeability***

Compound	**SMILES**	Permeability
Acebutolol	CCCC(=O)NC1=CC(=C(C=C1)OCC(CNC(C)C)O)C(=O)C	0
Aldosterone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4C(=O)CO)C=O)O	0
Antipyrine	CC1=CC(=O)N(N1C)C2=CC=CC=C2	1
Ascorbic acid	C(C(C1C(=C(C(=O)O1)O)O)O)O	0
Atenolol	CC(C)NCC(COC1=CC=C(C=C1)CC(=O)N)O	0
Atropine	CN1C2CCC1CC(C2)OC(=O)C(CO)C3=CC=CC=C3	0
Baclofen	C1=CC(=CC=C1C(CC(=O)O)CN)Cl	0
Bicuculline	CN1CCC2=CC3=C(C=C2C1C4C5=C(C6=C(C=C5)OCO6)C(=O)O4)OCO3	1
Bromperidol	C1CN(CCC1(C2=CC=C(C=C2)Br)O)CCCC(=O)C3=CC=C(C=C3)F	1
Bupropion	CC(C(=O)C1=CC(=CC=C1)Cl)NC(C)(C)C	1
Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	1
Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	1
Carbanilic acid	C1=CC=C(C=C1)NC(=O)O	1
Carebastine	CC(C)(C1=CC=C(C=C1)C(=O)CCCN2CCC(CC2)OC(C3=CC=CC=C3)C4=CC=CC=C4)C(=O)O	0
Cetirizine	C1CN(CCN1CCOCC(=O)O)C(C2=CC=CC=C2)C3=CC=C(C=C3)Cl	0
Chloramphenicol	C1=CC(=CC=C1C(C(CO)NC(=O)C(Cl)Cl)O)[N+](=O)[O-]	0
Chlorpromazine	CN(C)CCCN1C2=CC=CC=C2SC3=C1C=C(C=C3)Cl	1
Clonidine	C1CN=C(N1)NC2=C(C=CC=C2Cl)Cl	1
Codeine	CN1CCC23C4C1CC5=C2C(=C(C=C5)OC)OC3C(C=C4)O	1
Corticosterone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4C(=O)CO)C)O	1
Cytarabine	C1=CN(C(=O)N=C1N)C2C(C(C(O2)CO)O)O	0
Desipramine	CNCCCN1C2=CC=CC=C2CCC3=CC=CC=C31	1
Diacetylmorphine	CC(=O)OC1C=CC2C3CC4=C5C2(C1OC5=C(C=C4)OC(=O)C)CCN3C	1
Diazepam	CN1C(=O)CN=C(C2=C1C=CC(=C2)Cl)C3=CC=CC=C3	1
Dopamine	C1=CC(=C(C=C1CCN)O)O	0
Epinephrine	CNCC(C1=CC(=C(C=C1)O)O)O	0
Ethanol	CCO	1
Felbamate	C1=CC=C(C=C1)C(COC(=O)N)COC(=O)N	1
Fluphenazine	C1CN(CCN1CCCN2C3=CC=CC=C3SC4=C2C=C(C=C4)C(F)(F)F)CCO	1

Glutamine	C(CC(=O)N)C(C(=O)O)N	0
Haloperidol	C1CN(CCC1(C2=CC=C(C=C2)Cl)O)CCCC(=O)C3=CC=C(C=C3)F	1
Heptacaine	CCCCCCCOC1=CC=CC=C1NC(=O)OCC[NH+]2CCCCC2.[Cl-]	1
Histamine	C1=C(NC=N1)CCN	0
Hydrocortisone	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	0
Hydroxyzine	C1CN(CCN1CCOCCO)C(C2=CC=CC=C2)C3=CC=C(C=C3)Cl	1
Idoxuridine	C1C(C(OC1N2C=C(C(=O)NC2=O)I)CO)O	0
Imipramine	CN(C)CCCN1C2=CC=CC=C2CCC3=CC=CC=C31	1
Isopropyl alcohol	CC(C)O	1
Levomethadone	CCC(=O)C(CC(C)N(C)C)(C1=CC=CC=C1)C2=CC=CC=C2	1
Loperamide	CN(C)C(=O)C(CCN1CCC(CC1)(C2=CC=C(C=C2)Cl)O)(C3=CC=CC=C3)C4=CC=CC=C4	0
Mannitol	C(C(C(C(C(CO)O)O)O)O)O)O	0
Mescaline	COC1=CC(=CC(=C1OC)OC)CCN	0
Mesoridazine	CN1CCCCC1CCN2C3=CC=CC=C3SC4=C2C=C(C=C4)S(=O)C	0
Methotrexate	CN(CC1=CN=C2C(=N1)C(=NC(=N2)N)N)C3=CC=C(C=C3)C(=O)NC(CCC(=O)O)C(=O)O	0
Metoclopramide	CCN(CC)CCNC(=O)C1=CC(=C(C=C1OC)N)Cl	1
Metoprolol	CC(C)NCC(COC1=CC=C(C=C1)CCOC)O	1
Metrizamide	CC(=O)NC1=C(C(=C(C(=C1I)C(=O)NC2C(C(C(OC2O)CO)O)O)I)N(C)C(=O)C)I	1
Morphine	CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O	0
Nicotine	CN1CCCC1C2=CN=CC=C2	1
Norepinephrine	C1=CC(=C(C=C1C(CN)O)O)O	0
Paraldehyde	CC1OC(OC(O1)C)C	1
Pentylenetetrazol	C1CCC2=NN=NN2CC1	1
Perphenazine	C1CN(CCN1CCCN2C3=CC=CC=C3SC4=C2C=C(C=C4)Cl)CCO	1
Phenethylamine	C1=CC=C(C=C1)CCN	1
Phenobarbital	CCC1(C(=O)NC(=O)NC1=O)C2=CC=CC=C2	1
Phenytoin	C1=CC=C(C=C1)C2(C(=O)NC(=O)N2)C3=CC=CC=C3	1
Procaine	CCN(CC)CCOC(=O)C1=CC=C(C=C1)N	1
Promazine	CN(C)CCCN1C2=CC=CC=C2SC3=CC=CC=C31	1
Propranolol	CC(C)NCC(COC1=CC=CC2=CC=CC=C21)O	1

Protirelin	C1CC(N(C1)C(=O)C(CC2=CN=CN2)NC(=O)C3CCC(=O)N3)C(=O)N	1
Putrescine	C(CCN)CN	0
Ranitidine	CNC(=C[N+](=O)[O-])NCCSCC1=CC=C(O1)CN(C)C	0
Reserpine	COC1C(CC2CN3CCC4=C(C3CC2C1C(=O)OC)NC5=C4C=CC(=C5)OC)OC(=O)C6=CC(=C(C(=C6)OC)OC)OC	0
Rolipram	COC1=C(C=C(C=C1)C2CC(=O)NC2)OC3CCCC3	1
Roxatidine	C1CCN(CC1)CC2=CC(=CC=C2)OCCCNC(=O)CO	0
Serotonin	C1=CC2=C(C=C1O)C(=CN2)CCN	0
Sotalol	CC(C)NCC(C1=CC=C(C=C1)NS(=O)(=O)C)O	0
Spermidine	C(CCNCCCN)CN	0
Spermine	C(CCNCCCN)CNCCCN	0
Sucrose	C(C1C(C(C(C(O1)OC2(C(C(C(O2)CO)O)O)CO)O)O)O)O	0
Sulforidazine	CN1CCCCC1CCN2C3=CC=CC=C3SC4=C2C=C(C=C4)S(=O)(=O)C	1
Tamitinol	CCNCC1=C(C(=NC=C1CSC)C)O	1
Testosterone	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	1
Thioridazine	CN1CCCCC1CCN2C3=CC=CC=C3SC4=C2C=C(C=C4)SC	1
Triamterene	C1=CC=C(C=C1)C2=NC3=C(N=C(N=C3N=C2N)N)N	0
Trimelamol	CN(CO)C1=NC(=NC(=N1)N(C)CO)N(C)CO	0
Tryptamine	C1=CC=C2C(=C1)C(=CN2)CCN	1
Tryptophol	C1=CC=C2C(=C1)C(=CN2)CCO	1
Tyramine	C1=CC(=CC=C1CCN)O	0
Zonisamide	C1=CC=C2C(=C1)C(=NO2)CS(=O)(=O)N	1

- ***Table 2: Comparison of non-linear models performance***

Model	**R^2**	**MSE**	**RMSE**	**MAE**
Support Vector Regression (SVR)	0.8660	0.0001	0.0087	0.0034
Random Forest	0.9087	0.0001	0.0072	0.0055
XGBoost	0.9430	<0.0001	0.0057	0.0043
Neural Network	0.7459	0.0001	0.0120	0.0083