https://doi.org/10.48047/AFJBS.6.13.2024.6509-6524

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Development and validation of Z-Score-Based Machine Learning Method (ZBML) For Effective estimation of Drug Likeness

P.N.Shiammala[1] and N.Duraimutharasan2*

[1]Research Scholar, Department of Computer Science,

AMET deemed to be University, Kanathur, Chennai, and Tamilnadu, India

[2*]Professor, Department of Computer Science,

AMET deemed to be University, Kanathur, Chennai, and Tamilnadu, India

Volume 6, Issue 13, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

doi: 10.48047/AFJBS.6.13.2024.6509-6524

Abstract

Computational drug discovery plays a crucial role in identifying potential treatments for various diseases, particularly descriptor analysis has an emerged as a pivotal approach in the quest for novel therapeutics to combat skin diseases. By harnessing the power of computational algorithms researchers can rapidly sift through vast libraries of compounds, predicting their potential interactions with target proteins implicated in skin conditions. This targeted approach not only accelerates the drug discovery process but also enhances cost-efficiency by minimizing the need for laborious experimental validations. Computational methods facilitate the identification of compounds suited for skin penetration, target binding, and minimize toxicity by analyzing a range of molecular descriptors, including molecular weight, lipophilicity, and hydrogen bonding capacity. In this paper, we proposed a (ZBML) Z-Score based machine learning method to detect and remove the outliers in the collections of 13,241 small molecules dataset from the PubChem database and to perform Quantitative Structure-Activity Relationships (QSAR) for prediction of drug-likeness Lipinski descriptor. We made regression analysis of five types of machine learning algorithms out of those Linear models giving the best performance results for evaluation metrics of Mean Squared Error (MSE) = 1.10, Mean Absolute Error = 4.09, Root Mean Square Error = 1.049, R-squared (R2) coefficient (R2) =1.0.

Keywords: Drug Discovery, Machine Learning Algorithm, Z-Score Method, Lipinski descriptor.

1. Introduction

Molecular descriptors can be thought of as the "digital fingerprints" of molecules. They are quantitative representations that encapsulate information derived from the molecular structure, allowing this complex data to be analyzed statistically by ZBML method. [1] These descriptors make it possible to apply machine learning algorithms and other statistical tools to predict the behavior of molecules in various environments, which is invaluable in fields such as drug discovery, material science, and environmental science.[1] The fundamental purpose of molecular

descriptors is to convert chemical information into a form suitable for mathematical analysis and predictive modeling.

Molecular descriptors are numerical representations of chemical compounds that encode various physicochemical, topological, and structural properties. [2] These descriptors are crucial for computational chemistry and cheminformatics tasks such as quantitative structure-activity relationship (QSAR) modeling, virtual screening, and molecular similarity analysis.

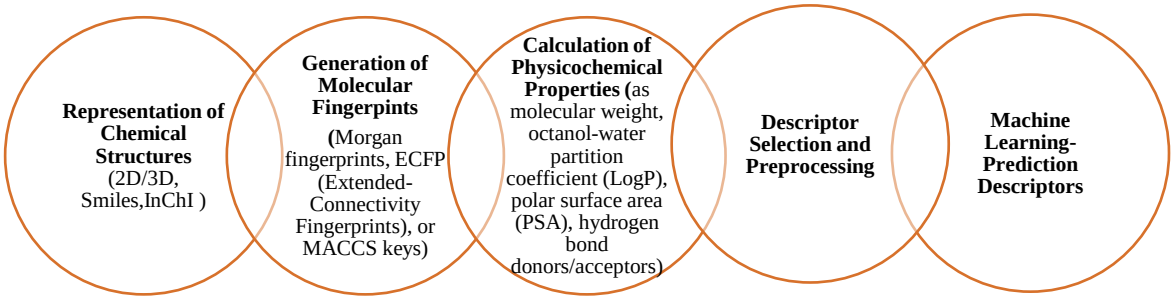

Figure 1Pipeline for extracting descriptors from the molecular structure

The figure 1 shows the pipeline for extract Lipinski's molecular descriptors from chemical structures involves several steps, including representation of data, data preprocessing, feature generation, and descriptor calculation. Step-1 we collect the data from the public database of the molecular information should be in the form of SMILES (Simplified Molecular Input Line Entry System) notation, InChI (International Chemical Identifier), or 2D/3D molecular structures. Step-2 using the computational package of RDkit to generate molecular fingerprints these fingerprints capture the presence or absence of specific substructures through a hashing algorithm and store them in a binary format. Where each bit represents the presence (1) or absence (0) of a particular substructure within the molecule. This vector is what is typically referred to as the Molecular Fingerprint. These fingerprints can now be used in machine learning models as features.Step-3 the fingerprint features is used to calculate the generation of physicochemical properties such as molecular weight, logP, hydrogen bond acceptor/donor, polar surface area and rotatable bonds. Step-4 After the successful extraction of Lipinski's descriptor ZBML method is used to pre-processing the data before given input to the machine learning algorithm. The ZBML based pre-processing could handle the data inconsistency, missing data to be handle with data imputation or removing etc. Data Exploration is an important step for every successful collection of data from

various resources. In this paper our proposed method is using Z-Score statistically approach for data exploration to evaluate the strict pre-processing method carried into the following steps.

1. Our model improve the performance for 100% predicting the Lipinski's molecular descriptor by Z-Score approach.

2. Z-Score approach handle data into two steps
    - Data analysis and visualization technique is applied to the prepared dataset
    - Visualization technique histogram plot is make to ensure the given dataset is adequate to applied Z-score method.
    - After the successful confirmation the calculation of Z-score is detecting and removing the outlier from the dataset along with Z-score value for each descriptors in the dataset. This technique is called as ZBML method.

"Lipinski's" descriptor is a well-known guideline in medicinal chemistry used to evaluate the drug-likeness of chemical compounds, particularly with regards to their potential for oral bioavailability. [3] Lipinski's Rule of Five states that, for a compound to be orally active, it should meet certain criteria regarding its physicochemical properties. These criteria are shown in the below table1

Table1 Explain the Threshold Limits for Lipinski's Descriptor

Criteria	**Threshold**	**Purpose**
Molecular weight	≤500 Dalton	Compounds heavier than 500 Daltons are less likely to be absorbed intestinally.
Number of hydrogen bond donors	≤ 5	A higher count of hydrogen bond donors typically decreases permeability across cell membranes.
Number of hydrogen bond acceptors	≤ 10	A higher count of hydrogen bond acceptors can affect the solubility and permeability.
LogP (octanol-water partition coefficient)	≤ 5	A measure of lipophilicity; values higher than 5 can imply poor solubility in water.
TPSA	≤ 140	Absorption and high likelihood of crossing the blood-brain barrier.

The idea behind Lipinski's Rule of Five is that compounds that violate one or more of these criteria may have difficulty crossing cell membranes or may exhibit poor oral bioavailability. However, it's important to note that Lipinski's Rule of Five is a guideline rather than a strict rule, and there are many successful drugs that do not adhere to these criteria.

2 Materials and Methods

2.1 Data Preparation

Collecting data from PubMed (https://pubmed.ncbi.nlm.nih.gov/) involves accessing the PubMed database, which is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. We extract the major causes of invasive skin related proteins and genes such as BRAF-3319, MEK-1176, MAPK-4001, MITF-882, CDK-184, NOTCH-92, and PTEN 2705. Total 13,241 small molecules were collected from the PubMed database. ZBML based pre-processing is performed on the collected dataset to check the null values and imputation method is used to fill the null values to prevent the data from loss and under fitting problems, then delete the irrelevant data, duplicate data and invalid smiles strings to prevent the data from overfitting problems. After the pre-processed dataset the final dataset is 7996 number of rows containing 2 columns such as id, smiles. Using python RDKit packages to extract the Lipinski's descriptors from the smiles structure of molecules.

2.2 Proposed Method

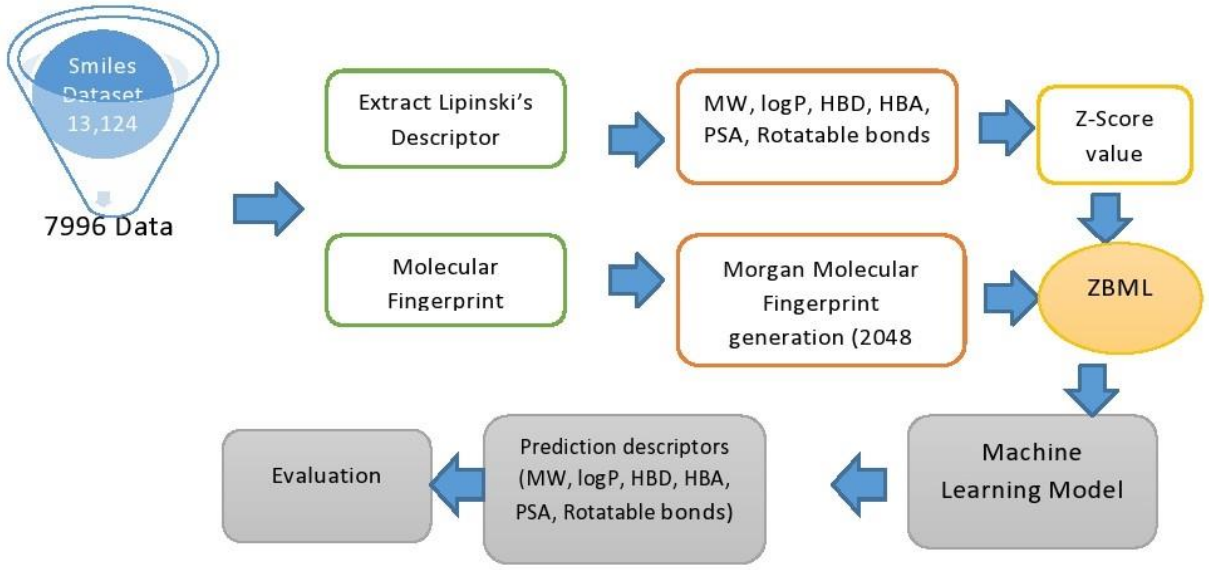

Figure 2 The Workflow of Proposed Z-Score Based Machine Learning Method (ZBML)

2.2.1 Z-Score Method

The Z-score method is a statistical technique used to standardize and compare data points from different distributions. It measures how many standard deviations a data point is from the mean of the distribution. The ZBML -score method is used to identify outliers and assess the significance of data points to improve the prediction accuracy of the human skin disease drug likeness properties figure 2.

The Z-score for a data point x in a dataset is calculated using the formula:

$$Z = x - \mu / \sigma$$

Where x is the data point, μ is the mean of the dataset, σ is the standard deviation of the dataset.

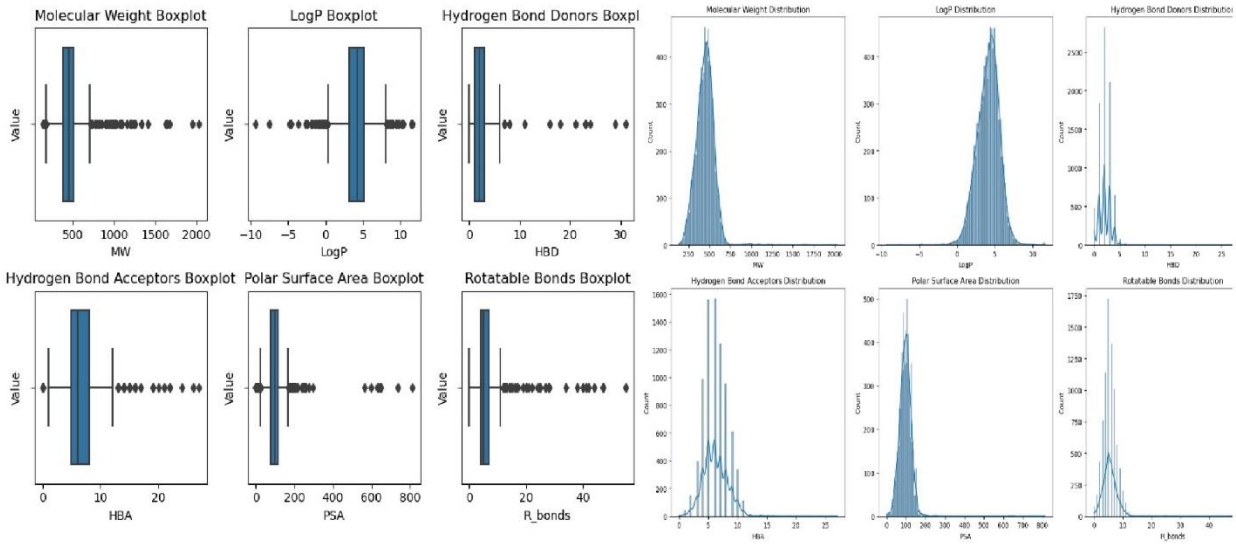

Figure 3 Box plot and Histogram visualization is used to detect outliers and data deviations in the dataset

Figure 3 Box plot describes the maximum numbers of data lies within the range of values such as molecular weight is 500 g/mol, LogP is 5, Hydrogen bond donor is 5, Hydrogen bond acceptor is 9, PSA is 175, Rotatable bonds is 15 out of these values all are outliers for the specified target of human skin disease drug like properties. For that we used the ZBML method to handle the outliers before input to the machine learning model. The below figure 4 explains the ranges of

value calculated after the ZBML method is used the total count of data is 7026 the mean value for MW-446, LogP-4.2, HBD-2.1, HBA-6.1, PSA-96.1, RB-5.5 and the data variation from the mean value is MW-87, LogP-1.2, HBD-1.0, HBA-1.8, PSA-24, RB-2.04. Using ZBML method to detect and remove the data deviation values then the result of data ranges for MW 223-672, LogP 1.17-7.06, HBD 0-4, HBA 2-10, PSA 29-163, RB 2-11. The percentages of drug likeness properties are 25% on 386, 3.33, 1.00, 5.00, 79.64, 4.00, 50% on 450, 4.30, 2.0, 6.0, 96.97, 5.0 and 75% on 507, 5.09, 3.0, 7.0, 112.98, and 7.0.

	MW	LogP	HBD	HBA	PSA	R_bonds	AR-c	MW_Zscore	LogP_Zscore
	HBD_Zscore	HBA_Zscore	PSA_Zscore	R_bonds_Zscore	AR-c_Zscore				
Count	7026.000000	7026.000000	7026.000000	7026.000000	7026.000000	7026.000000			
	7026.000000	7026.000000	7026.000000	7026.000000	7026.000000	7026.000000			
	7026.000000	7026.000000							
Mean	446.526647	4.202372	2.109166	6.169229	96.195253	5.506547			
	3.483917	-0.017178	0.058047	-0.025254	-0.027324	-0.029759			
	-0.020318	0.050695							
STD	87.903399	1.224911	1.006263	1.813582	24.738072	2.047973			
	0.854436	0.780960	0.830351	0.772297	0.854241	0.724996			
	0.731826	0.825835							
Min	223.659000	1.171200	0.000000	2.000000	29.100000	0.000000			
	2.000000	-1.997201	-1.996744	-1.644018	-1.991132	-1.996112			
	-1.988038	-1.383549							
25%	386.411000	3.333750	1.000000	5.000000	79.640000	4.000000			
	3.000000	-0.551263	-0.530780	-0.876528	-0.578059	-0.514942			
	-0.558671	-0.417023							
50%	450.542000	4.301150	2.000000	6.000000	96.970000	5.000000			
	4.000000	0.018496	0.125008	-0.109038	-0.107035	-0.007053			
	-0.201329	0.549503							

75%	507.598000	5.096900	3.000000	7.000000	112.985000	7.000000
	4.000000	0.525399	0.664436	0.658452	0.363989	0.462297
	0.513355	0.549503				
Max	672.812000	7.066400	4.000000	10.000000	163.740000	11.000000
	5.000000	1.993209	1.999534	1.425942	1.777062	1.949768
	1.942722	1.516029				

Figure 4 The Interpretation of Z-Scores approach in the given dataset

Figure 5 The Interpretation of Z-Scores approach in the given dataset total count is 7026 after removing outliers ,the mean value of molecular weight is 446 ,logP is 4, HBD is 2, HBA is 6,psa is 96 and rotatable bonds is 5 [25] . The proposed ZBML method results of Z –score value satisfies the drug-likeness property of Lipinski's rule. The value of z- score represent as 0 indicates that the data point is exactly at the mean. Positive Z-scores indicate values above the mean and Negative Z-scores indicate values below the mean. The Common thresholds for identifying outliers are Z-scores greater than +3 or less than -3. Our model used +2 and -2 threshold to detect outliers of the data point with Z-scores beyond a certain threshold. After removing the outliers from the dataset the z-score value representation of box plot and histogram.

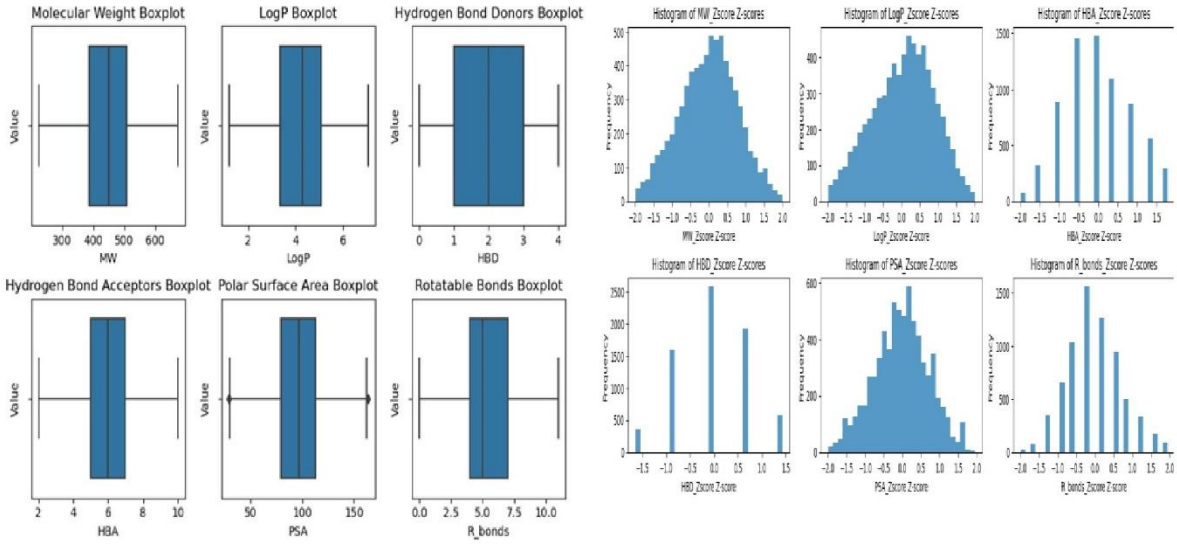

Figure 5 Box plot and Histogram visualization shows outliers detected dataset by ZBML method

2.2.2 Feature Generation

SMILES string of a molecule must be parsed and converted into a molecular 2D-structure representation that can be understood by computational tools RDKit. [4] This involves interpreting the SMILES notation to build a graph where nodes represent atoms and edges represent bonds. Once the SMILES is parsed, a molecular graph is generated. [4] This graph is a detailed representation of the molecule with all atoms and bonds specified as per the SMILES notation. From the molecular graph, various types of fingerprints can be calculated. [5] One popular type of fingerprint is the Molecular Fingerprint (MFP), often specifically referring to types like Extended Connectivity Fingerprints (ECFP) or Morgan Fingerprints in RDKit. [6] These fingerprints capture the presence or absence of specific substructures through a hashing algorithm and store them in a binary format.

The results from the hashing process are encoded into a binary vector, where each bit represents the presence (1) or absence (0) of a particular substructure within the molecule. [7] This vector is what is typically referred to as the Molecular Fingerprint. These fingerprints can now be used in machine learning models as features. [8] The following algorithm describes the steps involved in converting a molecular SMILES (Simplified Molecular Input Line Entry System) string into a Molecular Fingerprint (MFP).

Step1- SMILES string of a molecule Utilize RDKit MolFromSmiles function to parse the SMILES string and convert it into an RDKit molecule object.

Step2 - RDKit molecule object represents a molecular graph where each atom and bond from the SMILES is detailed in the object structure.

Step 3- Calculate Molecular Fingerprints by Extended Connectivity Fingerprints (ECFP), also known as Morgan Fingerprints (hashing 1024 0r 2048 bit).

Step 4- Use RDKit GetMorganFingerprintAsBitVect function to automate the conversion from hashed identifiers to a binary vector.

Step 5- Machine Learning use fingerprints as input features for predictive modelling to forecast properties or activities of molecules.

2.2.3 Machine Learning Models

2.2.3.1 Linear Regression Model

Linear regression stands is the statistical method for modeling the relationship between a dependent variable and one or more independent variables. Its premise rests on the assumption of a linear association between these variables, aiming to predict the dependent variable's value based on the independent ones. [9] In its simplest form, known as simple linear regression, there exists only one independent variable, while multiple linear regression extends this concept to encompass several predictors.[10] The model defines the relationship between the variables through a linear equation, typically represented as $Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\varepsilon$ where Y is the dependent variable, $X1,X2,\ldots,Xn$are the independent variables, $\beta 0,\beta 1,\ldots,\beta n$ are the coefficients representing the variables' effects, and $\varepsilon\varepsilon$ denotes the error term accounting for unexplained variance. The primary objective of linear regression is to estimate these coefficients in a manner that minimizes the discrepancy between the observed and predicted values of the dependent variable, typically achieved through the method of least squares. Linear regression finds widespread application across various domains, including economics, social sciences, engineering, and biology, serving purposes such as prediction, trend analysis, and hypothesis testing due to its simplicity, interpretability, and robustness when assumptions are met. [10] However, its reliance on the assumption of linearity and independence of errors underscores the importance of careful validation and consideration of model assumptions in practice.

2.2.3.2 Support Vector Machine Model

Support Vector Regression, or SVM regression, is used for solving regression problems. [11] Instead of predicting discrete class labels, SVR aims to predict continuous numeric values. The objective of SVR is to find a function that best fits the training data while limiting the deviation (epsilon) from the actual target values. In SVR, the training data points lying within the margin or on the wrong side of the margin are considered support vectors. [12] The distance between these support vectors and the regression function is minimized, while the deviation from the actual target values is controlled [30]. Support Vector Regression is a powerful algorithm in drug discovery that can effectively handle complex relationships between compound features and their continuous properties. By leveraging SVR, researchers can make quantitative predictions, gain insights into structure-property relationships, and guide compound optimization and prioritization efforts.

2.2.3.3 K-Nearest Neighbour Model

In regression, KNN predicts the value for a new data point by averaging the values of its k nearest neighbours. The average value serves as the predicted value for the new point. [13] [14] The approach can be extended to handle weighted averaging, where the neighbours' contributions are weighted based on their proximity to the new point.

2.2.3.4 Gradient Boosting Model

Gradient Boosting is a sophisticated ensemble learning technique renowned for its exceptional predictive performance across various machine learning tasks. At its core, Gradient Boosting builds a predictive model by sequentially combining multiple weak learners, typically decision trees, each focusing on the errors made by its predecessors.[15] This iterative process begins with an initial simple model, often a single decision tree or a constant value representing the average of the target variable. Subsequent models are then trained to predict the residuals, or errors, of the ensemble built so far. Through gradient descent optimization, these new models are fitted to minimize a chosen loss function, like mean squared error for regression or cross-entropy loss for classification, by adjusting their predictions in the direction that reduces the loss the most. [16] [17] The process continues, with each new model refining the predictions of the ensemble. Notably, Gradient Boosting allows for the integration of various weak learners and can adapt to complex datasets with nonlinear relationships. [16] While it excels in predictive accuracy and is robust to overfitting when appropriately regularized, Gradient Boosting can be computationally intensive and requires careful tuning of hyper parameters to achieve optimal performance. Popular implementations such as XGBoost, LightGBM, and CatBoost have made Gradient Boosting accessible and widely adopted in both academia and industry for its versatility and effectiveness in diverse real-world applications.

2.2.3.5 Random Forest Model

Random Forest Regression is used for solving regression problems. It aims to predict continuous numeric values instead of discrete class labels.[18] It combines the predictions of multiple decision trees to make accurate regression predictions. Each decision tree in the Random Forest estimates the output value based on a subset of the training data. The predictions from all the individual

regression trees are averaged to obtain the final regression prediction. [19] This aggregation helps to reduce the impact of individual tree biases and improve the overall prediction accuracy.

2.2.4 Evaluation Metrics

In evaluating the performance of our proposed regression model, which integrates z-score normalization with molecular fingerprinting for predicting Lipinski's descriptors, a robust evaluation framework is paramount. [20] [21] Lipinski's descriptors serve as fundamental indicators in drug design, influencing a compound's potential for oral bioavailability and pharmacokinetic properties. Our model's performance can be comprehensively evaluated through a suite of regression analysis metrics. [22] Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) offer valuable insights into the magnitude and distribution of prediction errors, crucial for understanding the model's precision in capturing Lipinski's descriptors. Furthermore, the Coefficient of Determination (*R*2 score) provides a holistic measure of the model's explanatory power, quantifying the proportion of variance in Lipinski's descriptors explained by our predictive features [23] [24].

Evaluating regression analysis for our proposed models involves several metrics to assess their performance.

1. Mean Absolute Error (MAE): This is the average of the absolute differences between predicted and actual values.

$$\text{ZBML of MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| = 4.08$$

2. Mean Squared Error (MSE): This metric squares the differences between predicted and actual values before averaging them.

$$\text{ZBML of MSE} = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y})^2 = 1.102$$

3. Root Mean Squared Error (RMSE): This is the square root of the MSE. It's in the same unit as the target variable and gives you an interpretable estimate of the average error.

$$\text{ZBML of RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2} = 1.04$$

4. Coefficient of Determination (R^2score): The R^2 metric, is a statistical measure used in regression analysis to assess the goodness of fit of a model. R^2 ranges from 0 to 1, Where the R^2 is 0 it does not explain any of the variability in the dependent variable, the R^2 is 1 the model explains all the variability in the dependent variable. It is useful for comparing different models. A higher R^2 indicates a better fit to the data.

$$\text{ZBML of } R^2 = \quad 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)2}{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y})2} = 1.0$$

3 .Results and Discussion

The comparative analysis of five regression models—linear regression, Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), and Random Forest (RF)—reveals that the linear regression model consistently outperforms the others across various evaluation metrics. Notably, the linear regression model achieved the lowest Mean Squared Error (MSE) and Mean Absolute Error (MAE), indicating it had the smallest average prediction errors. Additionally, it recorded the lowest Root Mean Square Error (RMSE), demonstrating its high precision and reliability. The R^2 coefficient for the linear model was the highest among the models, highlighting its strong ability to explain the variance in the dependent variable. Although SVM, KNN, GBM, and RF performed adequately, their error metrics were higher, and their R^2 values were lower than those of the linear regression model. This thorough evaluation underscores the linear regression model as the most effective and accurate choice for predicting Lipinski's descriptors using z-score normalized molecular fingerprints. The comparison results for each Lipinski's descriptor as shown as the figure 6.

Figure 7 explains residual plot is a graphical representation that shows the residuals on the vertical axis and the fitted values (or another variable) on the horizontal axis. Residuals are the differences between the observed and predicted values of the dependent variable. Horizontal Line at 0 interpretation means a residual plot where the residuals are concentrated along a horizontal line at 0 indicates that the differences between the observed and predicted values are minimal. In an ideal scenario, this means that the predicted values are almost perfectly matching the actual values for each data point.

MW Performance Comparison of Different Models Metrics

Model	Mean Squared Error	Mean Absolute Error	Root Mean Square Error	R-squared (R2) coefficient
linear	9.2031e-27	7.48798e-14	9.59328e-14	1
KNN	5182.88	57.7586	71.9922	0.2909
SVM	0.00573106	0.0601857	0.0757038	0.999999
GBM	0.610922	0.591603	0.781615	0.999916
RF	0.0223258	0.0640417	0.149418	0.999997

logP Performance Comparison of Different Models Metrics

Model	Mean Squared Error	Mean Absolute Error	Root Mean Square Error	R-squared (R2) coefficient
linear	4.48665e-30	1.6636e-15	2.11817e-15	1
KNN	0.358136	0.459074	0.598445	0.770431
SVM	0.00341538	0.0458312	0.0584412	0.997811
GBM	0.000116349	0.00816983	0.0107865	0.999925
RF	2.70266e-06	0.000819877	0.00164398	0.999998

HBD Performance Comparison of Different Models Metrics

Model	Mean Squared Error	Mean Absolute Error	Root Mean Square Error	R-squared (R2) coefficient
linear	1.45147e-28	9.49864e-15	1.20477e-14	1
KNN	0.166344	0.262731	0.407853	0.834265
SVM	0.00322298	0.0462223	0.0567713	0.996789
GBM	7.08201e-10	2.09977e-05	2.6612e-05	1
RF	0	0	0	1

HBA Performance Comparison of Different Models Metrics

Model	Mean Squared Error	Mean Absolute Error	Root Mean Square Error	R-squared (R2) coefficient
linear	1.25944e-29	2.75961e-15	3.54886e-15	1
KNN	0.632916	0.556899	0.79556	0.804868
SVM	0.00347016	0.046558	0.0589081	0.99893
GBM	3.40445e-09	4.29082e-05	5.83477e-05	1
RF	0	0	0	1

PSA Performance Comparison of Different Models Metrics

Model	Mean Squared Error	Mean Absolute Error	Root Mean Square Error	R-squared (R2) coefficient
linear	1.85166e-27	3.22802e-14	4.30309e-14	1
KNN	147.744	8.40241	12.155	0.749355
SVM	0.00558835	0.0579017	0.0747553	0.999991
GBM	0.041202	0.147366	0.202983	0.99993
RF	0.00161162	0.015666	0.040145	0.999997

RB Performance Comparison of Different Models Metrics

Model	Mean Squared Error	Mean Absolute Error	Root Mean Square Error	R-squared (R2) coefficient
linear	7.54531e-29	6.86759e-15	8.68637e-15	1
KNN	0.899915	0.635277	0.948638	0.77411
SVM	0.00424748	0.0499651	0.0651727	0.998934
GBM	6.36676e-09	5.31907e-05	7.9792e-05	1
RF	0	0	0	1

Figure 6 The results of comparative analysis for machine learning models by ZBML method

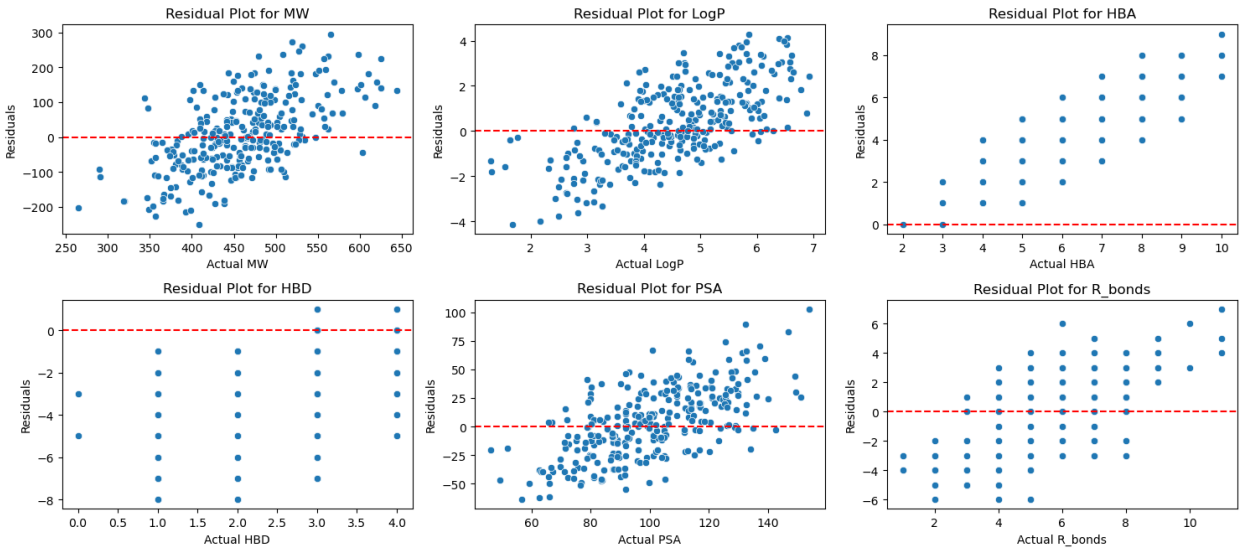

Figure 7 The Residual plot for predicted Lipinski's Descriptor

4. Conclusion

In conclusion, the proposed model ZBML Z-Score based machine learning method aimed to detect and remove outliers using z-score values, followed by a comparison of five different machine learning models: linear regression, support vector machines (SVM), k-nearest neighbours (KNN), gradient boosting, and random forest. Among these models, linear regression emerged as the top performer based on its performance metrics.

The utilization of z-score values for outlier detection and removal proved effective in enhancing the robustness and accuracy of the models by eliminating influential data points that could skew the results. This pre-processing step helped to improve the reliability and generalizability of the models by mitigating the impact of outliers on the training process.

While linear regression demonstrated the best performance in this study, it is important to acknowledge that the choice of the optimal model may vary depending on the specific dataset and problem domain. Therefore, further exploration and experimentation with alternative models and techniques could yield valuable insights and potentially improve overall model performance.

References.

1. T. Tran and C. Ekenna, (2022). "Molecular Descriptors Property Prediction via a Natural Language Processing Approach," IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 3492-3497.
2. Yang K, et al., (2019). Analyzing Learned Molecular Representations for Property Prediction. J Chem Inf Model. Aug 26;59(8):3370-3388.
3. Danishuddin, Khan AU.(2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today. 2016 Aug;21(8):1291-302.
4. Tuan Tran and Chinwe Ekenna, (2023). "Molecular Descriptors Property Prediction Using Transformer-Based Approach", International Journal of Molecular Sciences, vol.24, no.15, pp.11948.
5. Neves BJ,et al.,(2020). Deep Learning-driven research for drug discovery: Tackling Malaria. PLoS Comput Biol. 2020 Feb 18;16(2):e1007025.
6. Trinh, C. et al., (2023). On the Development of Descriptor-Based Machine Learning Models for Thermodynamic Properties: Part 1—From Data Collection to Model Construction: Understanding of the Methods and Their Effects. Processes 2023, 11, 3325.
7. Burden FR and Winkler DA. (1999). Robust QSAR models using Bayesian regularized neural networks. J Med Chem. Aug 12; 42(16):3183-7.
8. Heller S, et al., (2013). InChI - the worldwide chemical structure identifier standard. J Cheminform. 2013 Jan 24;5(1):7.
9. Shimakawa, H., Kumada, A. & Sato, (2024).M. Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. npj Comput Mater 10, 11 .
10. Wang J, et al., (2020).Cao D, Tang C, Chen X, Sun H, Hou T. Fast and accurate prediction of partial charges using Atom-Path-Descriptor-based machine learning. Bioinformatics. Sep 15;36(18):4721-4728.
11. Wen N, Liu G, Zhang J, Zhang R, Fu Y, Han X.(2022) A fingerprints based molecular property prediction method using the BERT model. J Cheminform. Oct 21;14(1):71.
12. Yang Q, Liu Y, Cheng J, Li Y, Liu S, Duan Y, Zhang L, Luo S.(2022). An Ensemble Structure and Physicochemical (SPOC) Descriptor for Machine-Learning Prediction of Chemical Reaction and Molecular Properties. Chemphyschem. 2022 Jul 19;23(14):e202200255.
13. Karami, Thomas & Hailu, Shumet & Feng, Shaoxin & Graham, Richard & Gukasyan, Hovhannes. (2021). Eyes on Lipinski's Rule of Five: A New “Rule of Thumb” for Physicochemical Design Space of Ophthalmic Drugs. Journal of Ocular Pharmacology and Therapeutics. 38. 10.1089/jop.2021.0069
14. Mishra H, Singh N, Lahiri T, Misra K.(2009) A comparative study on the molecular descriptors for predicting drug-likeness of small molecules. Bioinformation. 2009 Jun 13;3(9):384-8.

15. Xiujuan Liu,et al., (2022). Yueyue Shao, Tian Lu, Dongping Chang, Minjie Li, Wencong Lu, Accelerating the discovery of high-performance donor/acceptor pairs in photovoltaic materials via machine learning and density functional theory, Materials & Design, Volume 216 110561,ISSN 0264-1275.
16. Lohit, Niyatha & Singh, Ankit & Kumar, Adarsh & Singh, Harshwardhan & Yadav, Jagat Pal & Singh, Kuldeep & Kumar, Pradeep. (2023). Description and In silico ADME Studies of US-FDA Approved Drugs or Drugs under Clinical Trial which Violate the Lipinski's Rule of 5. Letters in Drug Design & Discovery. 20. 10.2174/1570180820666230224112505.
17. Bauer, C.A., Schneider, G. & Göller, A.H.(2019). Machine learning models for hydrogen bond donor and acceptor strengths using large and diverse training data generated by first-principles interaction free energies. J Cheminform 11, 59.
18. Kenney, D.H., Paffenroth, R.C., Timko, M.T. et al. (2023). Dimensionally reduced machine learning model for predicting single component octanol–water partition coefficients. J Cheminform 15, 9.
19. Yan-Kai Chen , Steven Shave * and Manfred Auer *.(2021) MRlogP: Transfer Learning Enables Accurate logP Prediction Using Small Experimental Training Datasets Processes, 9, 2029.
20. Win, Zaw-Myo & Cheong, Allen & Hopkins, W. (2023). Using Machine Learning To Predict Partition Coefficient (Log P) and Distribution Coefficient (Log D) with Molecular Descriptors and Liquid Chromatography Retention Time. Journal of Chemical Information and Modeling. 63. 10.1021/acs.jcim.2c01373.
21. Eelke B. Lenselink1 · Pieter F. W. Stouten1. (2021) . Multitask machine learning models for predicting lipophilicity (logP) in the SAMPL7 challenge.Journal of Computer-Aided Molecular Design 35:901–909.
22. Lowe Jr, Edward & Butkiewicz, Mariusz & Spellings, Matthew & Omlor, Albert & Meiler, Jens. (2011). Comparative analysis of machine learning techniques for the prediction of logP. 35-40. 10.1109/CIBCB.2011.5948478.
23. P. Maji and S. Paul, (2010)."Rough Sets for Selection of Molecular Descriptors to Predict Biological Activity of Molecules," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 639-648.
24. Mar-a Jimena Mart-nez, 1 Marina Razuc, 1, 2 and Ignacio Ponzoni.(2019). MoDeSuS: A Machine Learning Tool for Selection of Molecular Descriptors in QSAR Studies Applied to Molecular Informatics .Hindawi BioMed Research International, Article ID 2905203.
25. Khalid, A.; Badshah, G.; Ayub, N.; Shiraz, M.; Ghouse, M. Software Defect Prediction Analysis Using Machine Learning Techniques. Sustainability 2023, 15, 5517. https://doi.org/10.3390/su15065517