Taylor & Francis
Taylor & Francis Group

SAR and QSAR in Environmental Research

Predictive Modeling of Peroxisome Proliferator-activated Receptor Gamma (PPARγ) IC50 Inhibition by Emerging Pollutants Using Light Gradient Boosting Machine

Submission ID	247179946
Article Type	Original Article
Keywords	Modelling, LightGBM, Inhibition, Machine Learning, PPARγ, Organic pollutants
Authors	Adeboye Awomuti, Zhenyang Yu, Odunayo Blessing Adesina, Oluwarotimi Williams Samuel, ANNE MUMBI, Daqiang Yin

Predictive Modelling of Peroxisome Proliferator-activated Receptor Gamma (PPARγ) IC50 Inhibition by Emerging Pollutants Using Light Gradient Boosting Machine

A. Awomuti [a,b,c,*], Z. Yu [a,b,c], O. Adesina [a,c], O. W. Samuel [d,e], A. Mumbi [f,g], D. Yin [a,b]

a. State Key Laboratory of Pollution Control and Resource Reuse, Key Laboratory of Yangtze River Water Environment, Ministry of Education, College of Environmental Science and Engineering, Tongji University, Shanghai 200092, PR China

b. Shanghai Institute of Pollution Control and Ecological Security, Shanghai 200092, PR China

c. UNEP-Tongji Institute of Environment for Sustainable Development, College of Environmental Science and Engineering, Tongji University, Shanghai, China

d. School of Computing and Data Science Research Centre, University of Derby, Derby, DE22 3AW, United Kingdom

e. Faculty of Data Science and Information Technology, INTI International University, Nilai,71800, Malaysia

f. Department of Engineering, Harper Adams University, Edgmond, United Kingdom, TF10 8NB

g. Harper Adams Business School, Harper Adams University, Newport TF10 8NB, Shropshire, UK

Corresponding Author: Awomuti Adeboye [*]

College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China.

E-mail: adeboye@tongji.edu.cn; boyesky@yahoo.com

Predictive Modelling of Peroxisome Proliferator-activated Receptor Gamma (PPARγ) IC50 Inhibition by Emerging Pollutants Using Light Gradient Boosting Machine

Abstract

Peroxisome proliferator-activated receptor gamma (PPARγ), a critical nuclear receptor, plays a pivotal role in regulating metabolic and inflammatory processes. However, various environmental contaminants can disrupt PPARγ function, leading to adverse health effects. This study introduces a novel approach to predict the inhibitory activity (IC50 values) of 140 chemical compounds across 13 categories, including pesticides, organochlorines, dioxins, detergents, flame retardants, and preservatives, on PPARγ. The predictive model, based on the light-gradient boosting machine (LightGBM) algorithm, was trained on a dataset of 1804 molecules and demonstrated modest performance, with R2 scores of 0.82 and 0.59, Mean Absolute Error (MAE) of 0.38 and 0.58, and Root Mean Square Error (RMSE) of 0.54 and 0.76 for the training and test sets, respectively. This study provides novel insights into the interactions between emerging contaminants and PPARγ, highlighting the potential hazards and risks these chemicals may pose to public health and the environment. The ability to predict PPARγ inhibition by these hazardous contaminants demonstrates the value of this approach in guiding enhanced environmental toxicology research and risk assessment.

Keywords: Modelling; LightGBM; Inhibition; Machine Learning; PPARγ; Organic pollutants.

1. Introduction

Nuclear receptors play pivotal roles in regulating various critical biological processes [1]. Among the nuclear receptors, peroxisome proliferator-activated receptor gamma (PPARγ), as a transcription factor, regulates the metabolic balance, adipogenesis, lipid metabolism, inflammation, and cell differentiation [2,3]. Notably, nuclear receptors have essential functions in both disease development (e.g., obesity, diabetes and cancer) and also its treatment [4-7]. Endogenous (e.g., fatty acids) can bind to PPARγ and activate subsequent transcription and downstream processes [8-11]. Exogenous medicine (e.g., pioglitazone and rosiglitazone) agonize PPARγ and effectively treat metabolic diseases including type II diabetes by improving insulin sensitivity [12,13].

Unfortunately, nuclear receptors (e.g., PPARγ) also function as the targets of environmental pollutants to provoke health hazards. For example, bisphenol A and phthalates, possess PPARγ antagonistic activities, and their binding with the receptors can inhibit their transcription to disturb metabolic control and disease development [14-16]. Perfluoroalkyl substances (PFAS), one persistent toxic substance (PTS), are among the most widespread pollutants that have harmful effects on human health in relation to PPAR [17,18]. Novel industrial and household chemicals continue to emerge without sufficient toxicological scrutiny [19,20]. However, their effects on PPAR have not yet been systematically explored. Therefore, it is urgent to predict the effects of emerging pollutants on PPAR to facilitate earlier and faster screening and risk assessment [21].

Computational methods and large databases enable exploring molecular interactions at scales beyond laboratory studies [22,23]. Machine learning (ML) leverages bioactivity data on proteins to uncover patterns and develop predictive quantitative structure - activity relationship (QSAR) models [24-27]. Its efficiency speed up both toxicological and drug discoveries research to overcome the drawbacks of being slow and costly in traditional drug discovery [28-30]. Therefore, the combination of computational techniques with molecular responses will efficiently predict the potentials of environmental pollutants to interact with targets like PPARγ [31-34].

Computational approaches, such as quantitative structure-activity relationship (QSAR) modelling, have been employed to explore the interactions between environmental chemicals and the PPARγ receptor[35]. Previous QSAR studies have investigated the ability of various classes of compounds, including pesticides [36], pharmaceuticals [28], and industrial chemicals [37], to bind to and modulate the activity of PPARγ. These studies have demonstrated the utility of computational tools in predicting PPARγ ligand binding and identifying potential agonists or antagonists[38]. However, the development of robust and generalized QSAR models capable of accurately predicting PPARγ inhibition across a diverse set of emerging environmental pollutants remains an important research gap.

Among various computation methods, Light Gradient Boosting Machine (LightGBM) algorithm showed significant advantages[39]. LightGBM excels in processing large datasets efficiently, maintaining fast computational speeds and low memory usage, making it ideal for toxicity and drug discovery where rapid screening of thousands of molecules is required [40]. Its scalability and flexibility enable it to adapt to various chemical data and modelling scenarios [41]. Successfully used in bioinformatics and cheminformatics, LightGBM is trusted and credible for managing complex biological data, further justifying its selection for predicting median inhibition concentration (IC50) values in diverse chemical environments [42,43].

The present study established a LightGBM model with 1804, and achieved High R^2 scores, low MSE values, and competitive performance on other machine Learning evaluation metrics. The modelling process also demonstrated the importance of hyperparameter tuning in achieving optimal model performance. To further validate the robustness and predictive capabilities of our LightGBM model, our study extended the predictive model to encompass 140 chemical compounds across 13 distinct categories, including 19 pesticides, 19 organochlorine compounds, 15 common detergents and surfactants, 13 Preservatives, 10 Sweeteners and 10 dyes (Table 1 near here).

The model demonstrated comparable performance, predicting IC50 values that provided valuable insights into the potential interactions between these chemicals and PPARγ. In addition, the model provides novel insights into a range of other categories, including pigments, PFAS, PCBs, solvents, plastics, plasticizers, and more. Our study demonstrated the reliability and flexibility of our predictive LightGBM modelling method across a diverse range of chemical structures.

2. Materials and Methods

2.1. Dataset Extraction and Preprocessing

The present investigation leverages a comprehensive dataset capturing bioactivity measurements, comprising a substantial collection of experimental data points [44] using the Python RDKit library (version 2021.09.5) [45-49]. PPARγ ligands were selected based on their inhibition concentration (IC50) values, reported in μM units, resulting in a dataset of 1804 molecules **Table S1 (supplementary material)**[50] the initial dataset comprised 1887 molecules with reported PPARγ activity. After removing 83 molecules with missing data, the final dataset used for this study included 1804 unique PPARγ-active compounds. This preprocessing step ensured that the dataset was complete and suitable for the subsequent modelling and analysis tasks. IC50 values represent the concentration of ligand required to inhibit 50% of a target's activity, and are inversely related to binding affinity [51].

The dataset covers a wide range of chemical classes, including thiazolidinediones, oxazolidinones, and other heterocyclic scaffolds, ensuring a comprehensive representation of structural diversity. Additionally, the compounds exhibit a broad distribution of physicochemical properties, such as molecular weight, LogP, hydrogen bond donors/acceptors, which is crucial for developing a robust and generalizable predictive model.

The PPARγ activity data is represented by a variety of standard value types, providing a comprehensive assessment of the compounds' binding affinities. This diversity in both structural and activity data enables the model to capture the complex relationships between molecular features and PPARγ binding, enhancing its applicability to a wide range of drug discovery scenarios.

2.2. Descriptor Calculation and Fingerprint Generation

Molecular descriptors capturing drug-likeness were calculated using RDKit54 [52]. This included Lipinski descriptors, such as molecular weight (MW), calculated octanol-water partition coefficient (cLogP), number of hydrogen bond donors/acceptors, and number of rotatable bonds [53,54]. These descriptors provide essential information regarding the physicochemical properties of molecules [55,56]. **Table S2 (supplementary material)**

IC50 values in μM units extracted from the ChEMBL database, representing ligand concentrations required for 50% inhibition [57]. To facilitate comparison across compounds with diverse potencies on a logarithmic scale, IC50 values were transformed into pIC50 values using the negative decimal logarithm $\log_{10}(\mathrm{IC50})$as is commonly done in cheminformatics studies [58]. This logarithmic transformation standardizes potency measurements, making it simpler to compare the efficacy of different compounds. Lower pIC50 values indicate greater potency for easier modelling and evaluation versus direct concentration measurements.

The Predictive Activity of Drugs by Machine Learning (PADel) tool was used to encode each molecule as an 881-dimensional PubChem fingerprints, capturing structural information [59]. We utilized the PaDEL PubChem fingerprints due to their specific advantages in capturing molecular features relevant to biological activity. PaDEL PubChem fingerprints provide an extensive representation of molecular structures, incorporating a wide array of substructure information that enhances predictive modelling [42]. Compared to other fingerprint types, such as CDK, Extended, Estate, MACCS, and Klekota-Roth, PaDEL PubChem fingerprints offer a balanced trade-off between descriptor richness and computational efficiency, making them particularly suitable for large datasets

2.3. Dataset Splitting and Model Training

The calculated PADel PubChem molecular fingerprints were saved as a separate dataset, which was used to train and evaluate the regression model for PPARγ ligand prediction with IC50 values [60,61] **Table S3 (supplementary material)**. The dataset was randomly split into 80% for training (1443 molecules) and 20% for external testing (361 molecules) to evaluate the generalizability [62,63]. Normalization and handling of multicollinearity were automatically performed during the training process to ensure the model's robustness and accuracy [64,65].

2.4. Algorithm Evaluation and Selection

PyCaret (version 2.3) was used to conduct a comparative analysis with several other machine learning models, including Random Forest (RF), Gradient Boosting Regressor (GBR), Extreme Gradient Boosting (XGBoost), and others. Table 2 presents the performance metrics of 19 different algorithms, showcasing key indicators such as MAE, MSE, RMSE, and R^2[66,67]. LightGBM achieved an MAE of 0.5823 and an R^2 of 0.5999, positioning it among the top performers. In comparison, Random Forest yielded an MAE of 0.6136 and R^2 of 0.5637, while XGBoost delivered an MAE of 0.6402 and R^2 of 0.5205. Notably, both LightGBM and RF exhibited lower MAE values, indicating superior predictive accuracy over other algorithms, such as Gradient Boosting and Bayesian Ridge, which had higher MAEs of 0.6847 and 0.7009, respectively. (Table 2 near here)

Additionally, the computational efficiency of LightGBM is highlighted by its relatively low training time of 0.598 seconds, compared to Random Forest's 2.3213 seconds, making it a more suitable option for scenarios requiring rapid model training. This efficiency, coupled with competitive accuracy, underscores LightGBM's strengths in the context of PPARγ inhibition prediction[68-72]. PyCaret employs sensible defaults that are commonly accepted in the community for each algorithm, allowing for efficient hyperparameter tuning. During the model training process, PyCaret systematically evaluates a range of hyperparameters for each algorithm, optimizing their performance based on cross-validation metrics. The best hyperparameter values identified through this process are automatically logged, ensuring robust model selection. This approach not only streamlines the modelling workflow but also enhances the reliability of the results obtained.

We employed a robust 10-fold cross-validation approach to evaluate our predictive model's performance. This method is effective in assessing model generalizability while minimizing overfitting risks. The dataset was randomly divided into 10 equal-sized subsets, ensuring that each fold represented the overall dataset with a mixture of different chemical classes. Nine folds were combined to form the training set, while the remaining fold served as the testing set. This process was repeated until each fold had been used once as a testing set, yielding 10 performance estimates. For each iteration, we calculated key performance metrics, including R^2, MAE, MSE, and RMSE, which were averaged across all folds to provide a comprehensive assessment of model performance.

2.5. Training and Performance Evaluation of Light Gradient Boosting Machine (LightGBM) Model

The LightGBM model with 100 estimators and a learning-rate of 0.1 was selected based on cross-validation performance and tuning plots [73,74]. The model was trained on the PADel PubChem fingerprints and evaluated on external test data using the previously mentioned metrics [75]. The performance of the evaluated algorithms was assessed, and the LightGBM algorithm emerged as the best performing model. Various evaluation metrics were used to assess the model's performance[76-78]. The LightGBM model achieved R^2 scores of 0.82 for the training set and 0.59 for the testing set, indicating a modest fit for the data. Additionally, the model exhibited low MSE values of 0.29 and 0.58 on the training and testing sets, respectively, indicating accurate predictions[79,80]. The MAE scores were 0.38 and 0.58 for training and testing, respectively, further highlighting the precision of the model. The RMSE values for training and testing were 0.54 and 0.76, respectively, indicating a slight average deviation of the predictions from the actual values. The RMSLE scores of 0.07 for training and 0.10 for testing demonstrate the model's ability to accurately capture logarithmic errors. Furthermore, the MAPE scores of 0.06 for training and 0.09 for testing reflected the model's low average percentage error in prediction. These results show the capability of the LightGBM model in accurately predicting PPARg ligand activity based on IC50 values.

It also demonstrated the effectiveness of the LightGBM model in predicting PPARγ ligand activity based on IC50 values, as indicated by the R^2 score and low values for MSE, MAE, RMSE, RMSLE, and MAPE metrics. All model building and analyses were conducted in Python 3.9 using the mentioned libraries and their specified versions to ensure reproducibility[76,81,82].

3. Model establishment

3.1. Model evaluation metrics

LightGBM model’s performance was evaluated using various metrics for predicting PPAR ligand activity and IC50 values. The model showed promising results in multiple metrics, suggesting its potential as a reliable predictor of PPAR modulation[83].

MAE was used to assess prediction errors, representing the average absolute difference between predicted and actual values[60,84,85]. The LightGBM model displayed accurate predictions on both the training and testing datasets, with an MAE of 0.38 and 0.58, respectively (Table 3)[60,62,84]. The LightGBM model demonstrated a low MSE of 0.29 for the training dataset, indicating accurate predictions. The testing dataset also showed a competitive MSE of 0.58, validating the model's ability to predict PPAR ligand activity accurately.

Moreover, the Root Mean Squared Error (RMSE), a measure of the square root of the average squared differences between the predicted and actual values, was computed[86]. The LightGBM model had an RMSE of 0.54 on the training data, indicating the average error in predicted PPAR ligand activity. The model also had an RMSE of 0.76 on the testing data (Table 1), showing good generalization and reliable predictions[87,88]. The R^2 score measures the strength of the linear relationship between the independent and dependent variables, considering the proportion of variance explained by the model[89,90]. The LightGBM model achieved an impressive R^2 score of 0.82 on the training dataset, indicating that it can explain approximately 82% of the variance in PPAR ligand activity. Additionally, the model displayed a modest R^2 score of 0.59 on the testing dataset (Table 1), indicating a modest generalization to unseen data[91].

RMSLE accounts for logarithmic differences between predicted and actual values. LightGBM model's predictions aligned well with actual values with a training RMSLE of 0.07 and testing RMSLE of 0.10 (Table 1). The Mean Absolute Percentage Error (MAPE) metric, which measures percentage difference between predicted and actual values, was also calculated[92,93]. The LightGBM model demonstrated a low MAPE of 0.06 on the training dataset and 0.09 on the testing dataset (Table 3), suggesting that the model's predictions were relatively accurate compared to the actual values[94].

The LightGBM model showed Impressive performance in predicting PPAR ligand activity, with competitive results in various evaluation metrics, including R^2, MSE, MAE, RMSE, RMSLE, and MAPE. (Table 3 near here)

3.2. Improved optimal hyperparameter tuning for robust predictive modelling

The validation curve plot revealed valuable insights into the impact of the "max_depth" hyperparameter on model performance (Figure 1). The curve exhibited an initial increase in the R^2 score as the "max_depth" increased, indicating an improved performance.

Based on the validation curve, the optimal "max_depth" value was determined to be approximately 6.5, where the model achieved the highest R^2 score for the validation set. This value balances the complexity and generalization, ensuring that the model captures the relevant patterns without overfitting. (Figure 1 near here)

The validation curve plot also demonstrates a systematic approach to model hyperparameter tuning. This enhanced the reliability and generalizability of the predictive model. This highlights the importance of systematically evaluating the hyperparameter values to achieve the best performance and mitigate issues such as underfitting or overfitting, while fine-tuning the "max_depth" hyperparameter and optimizing the model's performance by leveraging the validation curve analysis.

3.3. Molecular descriptors relationships and feature selection

The heatmap illustrates the pairwise correlations among the variables in the dataset. MW has a moderately strong positive correlation with LogP and NumRings, suggesting that larger molecules tend to have higher LogP values and more rings (Figure 2). LogP also exhibits a moderate positive correlation with NumRings. Additionally, NumHDonors has weak positive correlations with MW and NumRings, while NumHAcceptors has weak negative correlations with LogP and strong positive correlations with MolSurfaceArea. (Figure 2 near here)

MolSurfaceArea has a weak positive correlation with NumHDonors and NumHAcceptors. This suggests that molecules with larger surface areas tend to contain more hydrogen donors and acceptors. pIC50 had a weak positive correlation with MW and LogP, indicating that larger molecules and those with higher partition coefficients may have higher pIC50 values. The correlation heatmap provides insights into the relationships between the variables in the dataset, thus providing a better understanding of how they are interrelated.

Feature importance analysis was performed to identify the most influential variables in predicting PPARγ inhibition. The SHAP (SHapley Additive exPlanations) values and permutation importance, which allowed us to quantify the contribution of each feature to the model's predictions were examined. The findings indicated that certain molecular characteristics, such as molecular weight, logP, and specific functional groups, significantly influenced inhibition activity. Discovering the impact of these features aided our understanding of the underlying mechanisms driving PPARγ inhibition, which could further facilitate research into targeted drug design and optimization. This information is crucial for both researchers and practitioners aiming to develop effective PPARγ inhibitors.

3.4. PPAR ligand characteristics through molecular descriptors correlation

The scatterplot (Figure 3) shows the relationship between the experimental (actual values) and predicted IC50 values (predicted values), enabling a comprehensive assessment of the accuracy of the predictive models. The graph demonstrates the correspondence between experimental and predicted values, and the regression line indicates the accuracy of the prediction. Variations from the line suggest potential inconsistencies or shortcomings in the predictive model. (Figure 3 near here)

The scatterplot provides several insights into predictive accuracy. A tight clustering of data points around the regression line suggests a high degree of agreement and reliability, indicating that the predictive model effectively captured the underlying relationships and displayed robust performance. Conversely, deviations from the regression line suggest potential systematic errors or limitations in the prediction models.

The regression line's slope and intercept offer insights into t he model's bias and precision. A slope close to 1 indicates a strong linear relationship, suggesting a reliable and unbiased predictive model. The scatterplot analysis and regression line reveal the pIC50 values' predictive accuracy.

4. Model application

4.1. Toxicological implications of studied compound exposure and comparative analysis of IC50 values and environmental detection

The toxicological implications of chemical exposure on PPARγ receptors are significant. Particularly, PFAS, pesticides and detergents display a wide range of effects on PPARγ [95]. The LightGBM model has been validated against known toxic compounds, confirming its accuracy in predicting PPARγ interaction. Notably, the model offers insights into the interaction potential of chemicals where empirical reports are not yet available, demonstrating its predictive power.

The IC50 values generated by the model are aligned with the detection levels of chemicals in environmental and biological samples. This is crucial for evaluating the actual risks associated with these chemicals. For example, Pesticides and Organochlorine Compounds often persist in soil and water, and their modeled IC50 values of 5.349 and 5.228, respectively, indicate a potential for PPARγ interaction that could be linked to observed health effects in populations exposed to these compounds.

It is important to note that the recorded IC50 values for these compounds in the PubChem database are 5.472 and 5.204, respectively. Similarly, the IC50 values obtained from the LightGBM model for Dioxins and Furans align closely with those from the study, indicating that the model accurately reflects the inhibitory potential of these toxicants. For Flame Retardants and Preservatives, there is a notable correlation between the modeled IC50 values and published reports linking these chemicals to PPAR-mediated diseases. This correlation underscores the validity of the model in identifying potential health risks.

4.2 Broader Environmental and Health Risk Discussion

The model's extensive coverage of various chemical c lasses, including Solvents, Plastics and Polymers, and Plasticizers, enables a comprehensive risk assessment. The IC50 values for Polychlorinated Biphenyls (PCBs) and other contaminants underscore their potential for PPARγ interaction and the consequent biological implications. By comparing these values, for instance, the average IC50 value for pesticides in a report is -1.905 [96]while the present study's predicted value is 5.349, similarly, the average IC50 values for preservatives in another research is -1.767 while the average IC50 value for preservatives in our study is 5.473. This study provides a nuanced understanding of the potential health risks associated with chronic exposure to these compounds.

The predictive model serves as an indispensable resource for enhancing our comprehension of PPARγ interactions across an array of chemical substances, thereby facilitating proactive risk assessment and the prioritization of compounds for additional toxicological scrutiny. The incorporation of a diverse range of chemicals in this study underscores the model's resilience and lays the groundwork for future research aimed at averting and mitigating chemical-induced diseases.

4.3 Enhanced Predictive Performance

The application of the LightGBM model in predicting IC50 values of 140 chemical compounds across 13 distinct categories (Table 1) was a critical test of the model's capability and generalizability. The list of tested 140 chemicals, smiles, structure, predicted IC50 and their calculated PADel-PubChem molecular fingerprints are presented in **Table S4 and S5 (supplementary material)** This was a significant step beyond the original model training, which was conducted using a dataset of 1,804 molecules. The performance of the model was assessed using similar evaluation metrics adopted in the training phase: R Square (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

The performance of the model on these new validation datasets was strikingly comparable to that of the training and test sets used to develop the model. The R^2 values indicates that the model could explain a substantial proportion of the variability in response data. Similarly, the MAE and RMSE values remained low, suggesting that the model predictions were close to the actual observations. Specifically, for pesticides, sweeteners, and preservatives, the model achieved an R^2 score above 0.81, demonstrating a strong correlation between the predicted and observed IC50 values. The MAE and RMSE were less than 0.40 and 0.56, respectively, indicating low prediction errors.

These results reinforce the robustness and reliability of our model in predicting PPARγ ligand activity based on IC50 values across different classes of chemicals. Furthermore, this underlines its potential as a versatile tool in environmental toxicology and pharmacology.

By accurately predicting the inhibitory effects of these chemicals on PPARγ, the LightGBM model can guide further laboratory investigations, help prioritize chemicals for regulatory scrutiny, and expedite the understanding of the potential toxicological implications of these chemicals.

4.4 Distribution of thirteen categories of Persistent Environmental Toxic Substances (PETS) Predicted IC50 Values

The distribution of predicted IC50 values for 13 chemical groups is illustrated in (Figure 4). The x-axis shows the predicted IC50 value, and the y-axis shows the frequency of that value, the values represent the potency of a chemical. The distribution of predicted IC50 values for the 13 chemical categories varies significantly. For instance, some categories, such as pesticides and organochlorine compounds, have a relatively high median value, suggesting that they are less likely to cause adverse effects. On the other hand, other categories, such as dioxins and furans, and flame retardants, have a relatively low median predicted IC50 value, indicating a higher likelihood of causing adverse effects. (Figure 4 near here)

This variety in predicted IC50 values implies that some chemicals within a category may be more potent and, thus, more likely to cause adverse effects than others. Identifying individual chemical potency is crucial when assessing their potential risks. The distribution of predicted IC50 values for the 13 chemical categories is a useful tool for understanding their potential hazards. By identifying the chemical category with the most potent and least potent chemicals and those with the widest and narrowest ranges of predicted IC50 values, it becomes possible to prioritize further examination for those posing significant risks.

The most potent chemical category is dioxins and furans, with a median predicted IC50 value of 4.5. The least potent chemical category is plastics and polymers, with a median predicted IC50 value of 6.3. The chemical category with the widest distribution of predicted IC50 values is preservatives, with a range of 4.4 to 6.8. The chemical category with the narrowest distribution of predicted IC50 values is plasticizers, with a range of 5.2 to 6.2.

Chemicals in categories with low median predicted IC50 values are more likely to cause adverse effects. Chemicals in categories with a wide range of predicted IC50 values may have varying potencies, with some being more likely to cause adverse effects than others. This distribution can help identify chemicals that are likely to cause adverse effects and prioritize them for further testing.

4.5 Exploring Variable Outliers and Complex Relationships

The pair plot in (Figure 5) displays the predicted IC50 values for different chemical categories: pesticides, organochlorine compounds, dioxins and furans, detergents and surfactants, and flame retardants. The results show that there are no outliers in the data and all of the data points are within the range of 4.5 to 6.0. However, some patterns are observed in the data. The predicted IC50 values for pesticides tend to be higher than the predicted IC50 values for the other chemical categories. The predicted IC50 values for detergents and surfactants tend to be lower than the predicted IC50 values for the other chemical categories. (Figure 5 near here)

The findings from the pair plots suggest that the potential risks posed by these chemicals may vary depending on the chemical category. Pesticides may pose a greater risk than the other chemical categories, while detergents and surfactants may pose a lower risk than other chemicals in this subset. The findings from the pair plots are consistent with the results of previous studies.

For example, a study by the Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Fuyang, China found that pesticides have a very high potential risk for human health[97].

The graph presented in (Figure 6) illustrates a pair plot of the projected IC50 values for a different set of chemicals with close correlations, including preservatives, pigments and dyes, PFAS, polychlorinated biphenyls (PCBs), and sweeteners. The diagonal components of the graph display histograms for each measurement, while the off-diagonal sections display scatterplots for each pair of measurements. (Figure 6 near here)

The histogram shows that the measurements are all normally distributed, except for the PCB-PIC50 measurement, which exhibits a slight deviation to the left. The scatterplots reveal that certain correlations exist between the measurements. For instance, the PRV-PIC50 measurement demonstrates a positive correlation with the Sweetners-PIC50 measurement, while the P&D-PIC50 measurement shows a negative correlation with both the PCB-PIC50 and PFAS-PIC50 measurements.

The pairplot shows that the five PIC50 measurements are all fairly normally distributed and that there are considerable correlations between them.

4.6 Analysis of Predicted IC50 Values of Thirteen Persistent Environmental Toxic Substances (PETS) by Category

The distribution of predicted IC50 values across various chemical categories is shown in (Figure 7). The boxplot depicts the median, 25th and 75th percentiles, and the whiskers extend to the most extreme values. The categories with the highest predicted IC50 values are plasticizers and P&P-PIC50, while the categories with the lowest predicted IC50 values are sweeteners and FR-PIC50. (Figure 7 near here)

This distribution exhibits statistically significant differences (Kruskal-Wallis test, $p < 0.05$). Further analysis using a post-hoc Dunn's test with a Bonferroni correction for multiple comparisons reveals that the following categories display significantly different predicted IC50 values: Plasticizers vs. Sweeteners ($p < 0.05$), Plasticizers vs. FR-PIC50 ($p < 0.05$), P&P-PIC50 vs. Sweeteners ($p < 0.05$), and P&P-PIC50 vs. FR-PIC50 ($p < 0.05$).

These findings indicate that the chemical category of a compound is a significant predictor of its predicted IC50 value. This information could be valuable in the development of new drugs and chemicals, as it could help identify compounds more likely to be active against a specific target e.g . The results of this analysis suggest that the LightGBM model developed in this study can predict IC50 values for different categories of PETS with reasonable accuracy. These models can be employed to prioritize chemicals for further testing and to assess the potential risks associated with these chemicals.

I need you to provide a direct manuscript input for the Limitations and Future Research section (Section 5: Limitations and Future Research 5.1 Limitations 5.2 Future Research Direction)

5. Limitations and Future Research

5.1 Limitations

While the Light Gradient Boosting Machine (LightGBM) model developed in this study shows promise in predicting PPARγ inhibition activity based on IC50 values, certain limitations must be acknowledged. First, the R^2 value of 0.59 for the test set indicates modest predictive performance and highlights the need for improvements that might have valuable contributions to the field. This discrepancy between the training set ($R^2 = 0.82$) and the test set suggests that the model may benefit from additional measures to address overfitting and improve its robustness.

Second, the dataset used in this study, though diverse, may still be limited in terms of size and representation. The test set contains structurally diverse compounds, and the inherent complexity of these molecules likely contributed to the variability in prediction accuracy. Furthermore, the dataset's relatively small size, particularly for the test set, may have constrained the model's ability to fully capture the intricate relationships between molecular features and inhibition activity.

Lastly, while the PADel-PubChem molecular fingerprints used in this study provide important information, the reliance on a single type of fingerprint representation may have limited the model's ability to fully account for the diverse physicochemical and structural properties of the chemical compounds. Exploring additional descriptor sets or hybrid approaches could enhance the model's overall predictive performance.

5.2 Future Research Direction

To address the limitations outlined above and further enhance the predictive capabilities of the model, future research could focus on several key areas:

1. **Dataset Expansion and Diversity:** Increasing the size and diversity of the dataset will be a priority to improve the model's generalizability. Incorporating additional chemical classes and expanding the dataset with experimentally validated IC50 values will help capture a wider range of structural and physicochemical properties, leading to better predictions.
2. **Advanced Modeling Techniques:** Future studies will explore ensemble modelling approaches, such as stacking or blending multiple algorithms, to leverage the strengths of different machine learning models. Additionally, the integration of deep learning architectures may help capture more complex relationships within the data.
3. **Feature Engineering and Descriptor Optimization:** To improve predictive performance, future research will evaluate alternative molecular descriptor sets, such as 3D structural descriptors or hybrid representations combining fingerprints with physicochemical properties. Feature selection techniques will also be employed to identify the most relevant features for predicting PPARγ inhibition activity.

4. **Incorporating Domain Knowledge:** Incorporating domain-specific knowledge, such as docking simulations or experimental binding affinities, may provide additional insights into the interactions between PPARγ and chemical compounds. This integration could improve the interpretability and accuracy of the predictions.

6. Conclusion

The study employs advanced computational methods and molecular investigations to unravel the complexities of PPARγ inhibitory activity. The Light Gradient Boosting Machine (LightGBM) algorithm excels in predicting PPARγ ligand activity based on IC50 values, as evidenced by high R2 scores, low Mean Squared Error (MSE), and strong performance in other metrics. Validation curve analysis and scatterplot analysis confirm the model's accuracy and reliability. The correlation heatmap reveals the relationships between molecular characteristics and their influence on ligand activity and inhibitory mechanisms. By combining computational expertise with ecological awareness, the study offers potential breakthroughs in drug discovery and toxicology for targeted treatments of metabolic disorders and inflammation. The findings provide a solid foundation for future research, deepening our understanding of molecular interactions, predictive modelling, and the practical implications of PPARγ modulation.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Zhang L, Sun W, Duan X, et al. Promoting differentiation and lipid metabolism are the primary effects for DINP exposure on 3T3-L1 preadipocytes [Article]. Environ Pollut. 2019;255.

[2] Garoche C, Boulahtouf A, Grimaldi M, et al. Interspecies Differences in Activation of Peroxisome Proliferator-Activated Receptor γ by Pharmaceutical and Environmental Chemicals. Environ Sci Technol. 2021;55(24):16489-16501.

[3] Virtue S, Petkevicius K, Maria Moreno-Navarrete J, et al. Peroxisome Proliferator-Activated Receptor gamma 2 Controls the Rate of Adipose Tissue Lipid Storage and Determines Metabolic Flexibility [Article]. Cell Reports. 2018;24(8):2005-+.

[4] Cao Y, Chen Y, Miao K, et al. PPAR gamma As a Potential Target for Adipogenesis Induced by Fine Particulate Matter in 3T3-L1 Preadipocytes [Article]. Environ Sci Technol. 2023;57(20):7684-7697.

[5] Villapol S. Roles of Peroxisome Proliferator-Activated Receptor Gamma on Brain and Peripheral Inflammation [Review]. Cellular and Molecular Neurobiology. 2018;38(1):121-132.

[6] Hall JM, Powell HA, Rajic L, et al. The Role of Dietary Phytoestrogens and the Nuclear Receptor PPAR gamma in Adipogenesis: An in Vitro Study [Article]. Environ Health Persp. 2019;127(3).

[7] Liu Z, Hua J, Cai W, et al. N-terminal truncated peroxisome proliferator-activated receptor-coactivator-1 alleviates phenylephrine-induced mitochondrial dysfunction and decreases lipid droplet accumulation in neonatal rat cardiomyocytes [Article]. Molecular Medicine Reports. 2018;18(2):2142-2152.

[8] Berger J, Moller DE. The mechanisms of action of PPARs. Annual Review of Medicine. 2002;53:409-435.

[9] Lehrke M, Lazar MA. The many faces of PPARγ [Review]. Cell. 2005;123(6):993-999.

[10] Ballav S, Biswas B, Sahu VK, et al. PPAR-gamma Partial Agonists in Disease-Fate Decision with Special Reference to Cancer [Review]. Cells. 2022;11(20).

[11] Zhang C, Zhang Y, Zhang C, et al. Pioglitazone increases VEGFR3 expression and promotes activation of M2 macrophages via the peroxisome proliferator-activated receptor [Article]. Molecular Medicine Reports. 2019;19(4):2740-2748.

[12] Liu CH, Lee TH, Lin YS, et al. Pioglitazone and PPAR-γ modulating treatment in hypertensive and type 2 diabetic patients after ischemic stroke: a national cohort study. Cardiovasc Diabetol. 2020;19(1):2.

[13] Gross B, Staels B. PPAR agonists: multimodal drugs for the treatment of type-2 diabetes. Best Practice & Research Clinical Endocrinology & Metabolism. 2007;21(4):687-710.

[14] Madsen MS, Broekema MF, Madsen MR, et al. PPAR gamma lipodystrophy mutants reveal intermolecular interactions required for enhancer activation [Article]. Nat Commun. 2022;13(1).

[15] Chang X, Shen Y, Yun L, et al. The antipsychotic drug olanzapine altered lipid metabolism in the common carp (Cyprinus carpio L.): Insight from the gut microbiota-SCFAs-liver axis [Article]. Sci Total Environ. 2023;856.

[16] Derangula M, Ruhinaz KK, Panati K, et al. Natural Product Ligands of the Peroxisome Proliferator-Activated Receptor Gamma as Anti-Inflammatory Mediators. Natural Products Journal. 2023;13(6):25-39.

[17] Christensen KY, Raymond M, Meiman J. Perfluoroalkyl substances and metabolic syndrome. Int J Hyg Envir Heal. 2019;222(1):147-153.

[18] Di Nisio A, Sabovic I, Valente U, et al. Endocrine Disruption of Androgenic Activity by Perfluoroalkyl Substances: Clinical and Experimental Evidence. J Clin Endocr Metab. 2019;104(4):1259-1271.

[19] Liyanage GY, Weerasekara MM, Manage PM. Screening and quantitative analysis of antibiotic resistance genes in hospital and aquaculture effluent in Sri Lanka as an emerging environmental contaminant. J Natl Sci Found Sri. 2022;50(2):361-370.

[20] Mostafa A, Shaaban H, Alqarni A, et al. Multi-class determination of pharmaceuticals as emerging contaminants in wastewater from Eastern Province, Saudi Arabia using eco-friendly SPE-UHPLC-MS/MS: Occurrence, removal and environmental risk assessment. Microchem J. 2023;187.

[21] El-Kalliny AS, Abdel-Wahed MS, El-Zahhar AA, et al. Nanomaterials: a review of emerging contaminants with potential health or environmental impact. Discov Nano. 2023;18(1).

[22] Hayden HL, Rochfort SJ, Ezernieks V, et al. Metabolomics approaches for the discrimination of disease suppressive soils for Rhizoctonia solani AG8 in cereal crops using H-1 NMR and LC-MS [Article]. Sci Total Environ. 2019;651:1627-1638.

[23] Shim J, Kim J, Noh H, et al. Robust, high-performance Cu-impregnated ZSM-5 zeolites for hydrocarbon removal during the cold-start period: Quantitative elucidation of effects of H+ and Na+ ions on performance and stability. Chemical Engineering Journal. 2024;494:153258.

[24] Bart S, Jager T, Robinson A, et al. Predicting Mixture Effects over Time with Toxicokinetic-Toxicodynamic Models (GUTS): Assumptions, Experimental Testing, and Predictive Power [Article]. Environ Sci Technol. 2021;55(4):2430-2439.

[25] Dusserre C, Mollergues J, Lo Piparo E, et al. Using bisphenol A and its analogs to address the feasibility and usefulness of the CALUX-PPAR gamma assay to identify chemicals with obesogenic potential [Article]. Toxicology in Vitro. 2018;53:208-221.

[26] Zhang G-L, Liu F-Y, Zhang J, et al. Integrated in silico-in vitro screening of ovarian cancer peroxisome proliferator-activated receptor-gamma agonists against a biogenic compound library [Article]. Medicinal Chemistry Research. 2018;27(1):341-349.

[27] Vitale CM, Di Guardo A. A review of the predictive models estimating association of neutral and ionizable organic chemicals with dissolved organic carbon [; Review]. The Science of the total environment. 2019;666:1022-1032.

[28] Xia J, Yang L, Dong L, et al. Cefminox, a Dual Agonist of Prostacyclin Receptor and Peroxisome Proliferator-Activated Receptor-Gamma Identified by Virtual Screening, Has Therapeutic Efficacy against Hypoxia-Induced Pulmonary Hypertension in Rats [Article]. Frontiers in Pharmacology. 2018;9.

[29] Li C-H, Ren X-M, Ruan T, et al. Chlorinated Polyfluorinated Ether Sulfonates Exhibit Higher Activity toward Peroxisome Proliferator-Activated Receptors Signaling Pathways than Perfluorooctanesulfonate [Article]. Environ Sci Technol. 2018;52(5):3232-3239.

[30] Zhong S, Zhang Y, Zhang H. Machine Learning-Assisted QSAR Models on Contaminant Reactivity Toward Four Oxidants: Combining Small Data Sets and Knowledge Transfer [Article]. Environ Sci Technol. 2022;56(1):681-692.

[31] Hong F, Xu P, Zhai Y. The Opportunities and Challenges of Peroxisome Proliferator-Activated Receptors Ligands in Clinical Drug Discovery and Development [Review]. International Journal of Molecular Sciences. 2018;19(8).

[32] Capitao AMF, Lopes-Marques MS, Ishii Y, et al. Evolutionary Exploitation of Vertebrate Peroxisome Proliferator-Activated Receptor gamma by Organotins [Article]. Environ Sci Technol. 2018;52(23):13951-13959.

[33] Cazzaniga A, Locatelli L, Castiglioni S, et al. The Contribution of EDF1 to PPAR gamma Transcriptional Activation in VEGF-Treated Human Endothelial Cells [Article]. International Journal of Molecular Sciences. 2018;19(7).

[34] Heudobler D, Rechenmacher M, Lueke F, et al. Peroxisome Proliferator-Activated Receptors (PPAR) Agonists as Master Modulators of Tumor Tissue [Review]. International Journal of Molecular Sciences. 2018;19(11).

[35] Ardenkjær-Skinnerup J, Nissen ACVE, Nikolov NG, et al. Orthogonal assay and QSAR modelling of Tox21 PPARγ antagonist in vitro high-throughput screening assay. Environmental Toxicology and Pharmacology. 2024;105:104347.

[36] Zhang G-L, Liu F-Y, Zhang J, et al. Integrated in silico–in vitro screening of ovarian cancer peroxisome proliferator-activated receptor-γ agonists against a biogenic compound library. Medicinal Chemistry Research. 2018;27(1):341-349.

[37] Hong F, Xu P, Zhai Y. The Opportunities and Challenges of Peroxisome Proliferator-Activated Receptors Ligands in Clinical Drug Discovery and Development. International Journal of Molecular Sciences. 2018;19(8):2189.

[38] Guasch L, Sala E, Valls C, et al. Development of docking-based 3D-QSAR models for PPARgamma full agonists. J Mol Graph Model. 2012;36:1-9.

[39] Pan Z, Lu W, Wang H, et al. Groundwater contaminant source identification based on an ensemble learning search framework associated with an auto xgboost surrogate. Environ Modell Softw. 2023;159:105588.

[40] Abdulhakeem Mansour Alhasbary A, Hashimah Ahamed Hassain Malim N. Turbo Similarity Searching: Effect of Partial Ranking and Fusion Rules on ChEMBL Database [Article]. Molecular Informatics. 2022;41(5).

[41] Dhar J. An adaptive intelligent diagnostic system to predict early stage of parkinson's disease using two-stage dimension reduction with genetically optimized lightgbm algorithm. Neural Comput Appl. 2022;34(6):4567-4593.

[42] Gao HS, Ye ZYF, Dong J, et al. Predicting drug/phospholipid complexation by the lightGBM method. Chem Phys Lett. 2020;747.

[43] Han L, Zhu YZT, Chen YW, et al. A LightGBM and XGBoost Learning Method for Postoperative Critical Illness Key Indicators Analysis. Ksii T Internet Inf. 2023;17(8):2016-2029.

[44] Chembl PPARg IC50 Bioactivity Dataset [Internet]. UK: European Molecular Biology Laboratory. 2022 [cited February 2024]. Available from: https://tinyurl.com/23fezxph.

[45] Pundir H, Joshi T, Pant M, et al. Identification of SARS-CoV-2 RNA dependent RNA polymerase inhibitors using pharmacophore modelling, molecular docking and molecular dynamics simulation approaches [Article]. Journal of Biomolecular Structure and Dynamics. 2022;40(24):13366-13377.

[46] Bort W, Mazitov D, Horvath D, et al. Inverse QSAR: Reversing Descriptor-Driven Prediction Pipeline Using Attention-Based Conditional Variational Autoencoder [Article]. Journal of Chemical Information and Modeling. 2022;62(22):5471-5484.

[47] Ebejer JP, Charlton MH, Finn PW. Are the physicochemical properties of antibacterial compounds really different from other drugs? [Article]. Journal of Cheminformatics. 2016;8(1).

[48] Li R, Bajorath J. Systematic assessment of scaffold distances in ChEMBL: Prioritization of compound data sets for scaffold hopping analysis in virtual screening [Article]. Journal of Computer-Aided Molecular Design. 2012;26(10):1101-1109.

[49] Warr WA. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI) [Article]. Journal of Computer-Aided Molecular Design. 2009;23(4):195-198.

[50] Wang Y, Wang T. Application of Improved LightGBM Model in Blood Glucose Prediction. Appl Sci-Basel. 2020;10(9).

[51] Goebel T, Diehl O, Heering J, et al. Zafirlukast Is a Dual Modulator of Human Soluble Epoxide Hydrolase and Peroxisome Proliferator-Activated Receptor gamma [Article]. Frontiers in Pharmacology. 2019;10.

[52] Hu XH, Yao ZY. Selection of Outline Descriptors Based on LightGBM with Application to Infrared Image Target Recognition. Sci Programming-Neth. 2021;2021.

[53] Alakhdar AA, Saleh AH, Arafa RK. Targeting homologous recombination (HR) repair mechanism for cancer treatment: discovery of new potential UCHL-3 inhibitors via virtual screening, molecular dynamics and binding mode analysis [Article]. Journal of Biomolecular Structure and Dynamics. 2022;40(1):276-289.

[54] Zhu QX, Zhang N, He YL, et al. Novel Imbalanced Fault Diagnosis Method based on CSMOTE integrated with LSDA and LightGBM for Industrial Process. Int C Control Decisi. 2022:326-331.

[55] Shovan SM, Hasan MA, Islam MR. Accurate Prediction of Formylation PTM Site using Multiple Feature Fusion with LightGBM Resolving Data Imbalance Issue. Int Conf Comput Info. 2020.

[56] Wang MH, Yue LL, Yang XH, et al. Fertility-LightGBM: A fertility-related protein prediction model by multi-information fusion and light gradient boosting machine. Biomed Signal Proces. 2021;68.

[57] Nagar N, Saxena H, Pathak A, et al. A review on structural mechanisms of protein-persistent organic pollutant (POP) interactions. Chemosphere. 2023;332.

[58] Yao XT, Fu XL, Zong CF. Short-Term Load Forecasting Method Based on Feature Preference Strategy and LightGBM-XGboost. Ieee Access. 2022;10:75257-75268.

[59] Zhang J, Mucs D, Norinder U, et al. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. Journal of Chemical Information and Modeling. 2019;59(10):4150-4158.

[60] Mehtab V, Alam S, Povari S, et al. Reduced Order Machine Learning Models for Accurate Prediction of CO2 Capture in Physical Solvents [Article; Early Access]. Environ Sci Technol. 2023.

[61] O'Boyle NM. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI [Article]. Journal of Cheminformatics. 2012;4(9).

[62] Hong N, Hama T, Suenaga Y, et al. Application of a modified conceptual rainfall-runoff model to simulation of groundwater level in an undefined watershed [Article]. Sci Total Environ. 2016;541:383-390.

[63] Zhang DY, Gong YC. The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure. Ieee Access. 2020;8:220990-221003.

[64] Liu LJ, Wang L, Yu Z. Remaining Useful Life Estimation of Aircraft Engines Based on Deep Convolution Neural Network and LightGBM Combination Model. Int J Comput Int Sys. 2021;14(1).

[65] Awomuti A, Alimo PK, Lartey-Young G, et al. Towards adequate policy enhancement: An AI-driven decision tree model for efficient recognition and classification of EPA status via multi-emission parameters. City and Environment Interactions. 2023;20:100127.

[66] Schaduangrat N, Anuwongcharoen N, Charoenkwan P, et al. DeepAR: a novel deep learning-based hybrid framework for the interpretable prediction of androgen receptor antagonists [Article]. Journal of Cheminformatics. 2023;15(1).

[67] Zhao Q, Yu Y, Gao Y, et al. Machine Learning-Based Models with High Accuracy and Broad Applicability Domains for Screening PMT/vPvM Substances [Article]. Environ Sci Technol. 2022;56(24):17880-17889.

[68] Yang HZ, Chen ZJ, Yang HJ, et al. Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison. Ieee Access. 2023;11:23366-23380.

[69] Zhao GS, Wang Y, Wang J. Intrusion Detection Model of Internet of Things Based on LightGBM. Ieice T Commun. 2023;E106b(8):622-634.

[70] Hatmal MM, Jaber S, Taha MO. Combining molecular dynamics simulation and ligand-receptor contacts analysis as a new approach for pharmacophore modeling: beta-secretase 1 and check point kinase 1 as case studies [Article]. Journal of Computer-Aided Molecular Design. 2016;30(12):1149-1163.

[71] Shehadeh A, Alshboul O, Al Mamlook RE, et al. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automat Constr. 2021;129.

[72] Saber M, Boulmaiz T, Guermoui M, et al. Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. Geocarto Int. 2022;37(25):7462-7487.

[73] Joudaki A, Takeda J, Masuda A, et al. FexSplice: A LightGBM-Based Model for Predicting the Splicing Effect of a Single Nucleotide Variant Affecting the First Nucleotide G of an Exon. Genes-Basel. 2023;14(9).

[74] Liljestrand D, Johnson R, Skiles SM, et al. Quantifying regional variability of machine-learning-based snow water equivalent estimates across the Western United States. Environ Modell Softw. 2024;177:106053.

[75] Zhang HQ, Li YQ. LightGBM Indoor Positioning Method Based on Merged Wi-Fi and Image Fingerprints. Sensors-Basel. 2021;21(11).

[76] Talkhi N, Nooghabi MJ, Esmaily H, et al. Prediction of serum anti-HSP27 antibody titers changes using a light gradient boosting machine (LightGBM) technique. Sci Rep-Uk. 2023;13(1).

[77] Niazkar M, Menapace A, Brentan B, et al. Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). Environ Modell Softw. 2024;174:105971.

[78] Thongthammachart T, Araki S, Shimadera H, et al. Incorporating Light Gradient Boosting Machine to land use regression model for estimating NO2 and PM2.5 levels in Kansai region, Japan. Environ Modell Softw. 2022;155:105447.

[79] Gan M, Pan SQ, Chen YP, et al. Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River. J Mar Sci Eng. 2021;9(5).

[80] Hajihosseinlou M, Maghsoudi A, Ghezelbash R. A Novel Scheme for Mapping of MVT-Type Pb-Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm. Nat Resour Res. 2023.

[81] Candido C, Blanco AC, Medina J, et al. Improving the consistency of multi-temporal land cover mapping of Laguna lake watershed using light gradient boosting machine

(LightGBM) approach, change detection analysis, and Markov chain. Remote Sens Appl. 2021;23.

[82] Joshi J. Chapter 9 - Python, a reliable programming language for chemoinformatics and bioinformatics. In: Sharma N, Ojha H, Raghav PK, et al., editors. Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences: Academic Press; 2021. p. 279-304.

[83] Cortes-Ciriano I. Benchmarking the Predictive Power of Ligand Efficiency Indices in QSAR [Article]. Journal of Chemical Information and Modeling. 2016;56(8):1576-1587.

[84] Cai J, Zheng P, Qaisar M, et al. Prediction and quantifying parameter importance in simultaneous anaerobic sulfide and nitrate removal process using artificial neural network [Article]. Environ Sci Pollut R. 2015;22(11):8272-8279.

[85] Chellali MR, Abderrahim H, Hamou A, et al. Artificial neural network models for prediction of daily fine particulate matter concentrations in Algiers [Article]. Environ Sci Pollut R. 2016;23(14):14008-14017.

[86] Endo S, Hammer J, Matsuzawa S. Experimental Determination of Air/Water Partition Coefficients for 21 Per- and Polyfluoroalkyl Substances Reveals Variable Performance of Property Prediction Models [Article]. Environ Sci Technol. 2023;57(22):8406-8413.

[87] Okoji AI, Anozie AN, Omoleye JA, et al. Evaluation of adaptive neuro-fuzzy inference system-genetic algorithm in the prediction and optimization of NOx emission in cement precalcining kiln [Article; Early Access]. Environ Sci Pollut R. 2023.

[88] Choi H, Zdeb M, Perera F, et al. Estimation of chronic personal exposure to airborne polycyclic aromatic hydrocarbons [Article]. Sci Total Environ. 2015;527:252-261.

[89] Liao K, Wang Q, Wang S, et al. Bayesian Inference Approach to Quantify Primary and Secondary Organic Carbon in Fine Particulate Matter Using Major Species Measurements [Article]. Environ Sci Technol. 2023;57(13):5169-5179.

[90] Omeka ME. Evaluation and prediction of irrigation water quality of an agricultural district, SE Nigeria: an integrated heuristic GIS-based and machine learning approach [Article; Early Access]. Environ Sci Pollut R. 2023.

[91] Soni K, Parmar KS, Kapoor S, et al. Statistical variability comparison in MODIS and AERONET derived aerosol optical depth over Indo-Gangetic Plains using time series modeling [Article]. Sci Total Environ. 2016;553:258-265.

[92] Zhang B, Ling L, Zeng L, et al. Multi-step prediction of carbon emissions based on a secondary decomposition framework coupled with stacking ensemble strategy [Article]. Environ Sci Pollut R. 2023;30(27):71063-71087.

[93] Sappl J, Harders M, Rauch W. Machine learning for quantile regression of biogas production rates in anaerobic digesters [Article]. Sci Total Environ. 2023;872.

[94] Shi G-L, Tian Y-Z, Ye S, et al. Source apportionment of synchronously size segregated fine and coarse particulate matter, using an improved three-way factor analysis model [Article]. Sci Total Environ. 2015;505:1182-1190.

[95] Tokranov AK, Nishizawa N, Amadei CA, et al. How Do We Measure Poly- and Perfluoroalkyl Substances (PFASs) at the Surface of Consumer Products? Environ Sci Tech Let. 2019;6(1):38-43.

[96] Ma M, Chen C, Yang G, et al. Combined cytotoxic effects of pesticide mixtures present in the Chinese diet on human hepatocarcinoma cell line. Chemosphere. 2016;159:256-266.

[97] Liu Y, Shen D, Li S, et al. Residue levels and risk assessment of pesticides in nuts of China. Chemosphere. 2016;144:645-651.

Figure Captions

- **Figure 1**. Model validation curve plot demonstrating Optimal Hyperparameter Tuning for Robust Predictive Modelling

- **Figure 2**. The Descriptor Correlation heatmap reveals the interactions between various chemical descriptors

- **Figure 3**. Model correlation plot, shows a well-fitted Algorithm

- **Figure 4**. Distribution of Predicted IC50 Values Amongst the 13 Chemicals

- **Figure 5**. Predicted IC50 Value Pairplot Variable Exploration for chemical categories: pesticides, organochlorine compounds, dioxins and furans, detergents and surfactants, and flame retardants.. PST= pesticides, OCC= organochlorine compounds, D&F= dioxins and furans, D&S= detergents and surfactants, and FR= flame retardants

- **Figure 6**. Predicted IC50 Value Pairplot Variable Exploration for chemical categories: preservatives, pigments and dyes, per- and polyfluoroalkyl substances (PFAS), polychlorinated biphenyls (PCBs), and sweeteners. PRV= preservatives, P&D= pigments and dyes

- **Figure 7**. Boxplot Analysis of Predicted IC50 Values by Chemical Category: pesticides, organochlorine compounds, dioxins and furans, detergents and surfactants, and flame retardants. PST= pesticides, OCC= organochlorine compounds, D&F= dioxins and furans

Table 1. Summary of the 13 chemical categories and the number of compounds included in the study. These categories encompass a wide range of environmental contaminants, including pesticides, organochlorines, dioxins, flame retardants, and others, tested for their potential to inhibit PPARγ

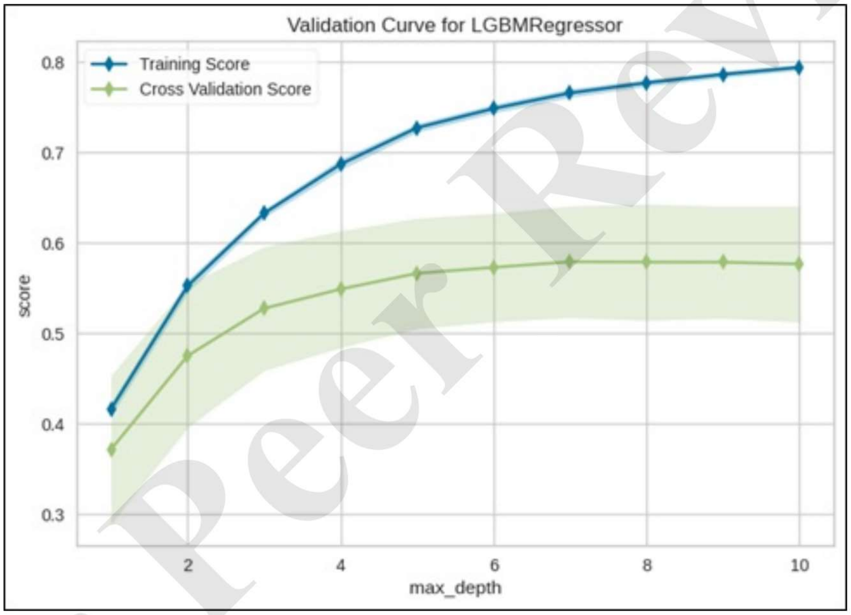
Chemical Category	Sample
Pesticides	19
Organochlorine Compounds	19
Dioxins and Furans	17
Detergents and Surfactants	15
Flame Retardants	14
Preservatives	13
Sweeteners	10
Pigments and Dyes	10
PFAS	9
Polychlorinated Biphenyls (PBCs)	8
Solvents	3
Plastics and Polymers	2
Plasticizers	1

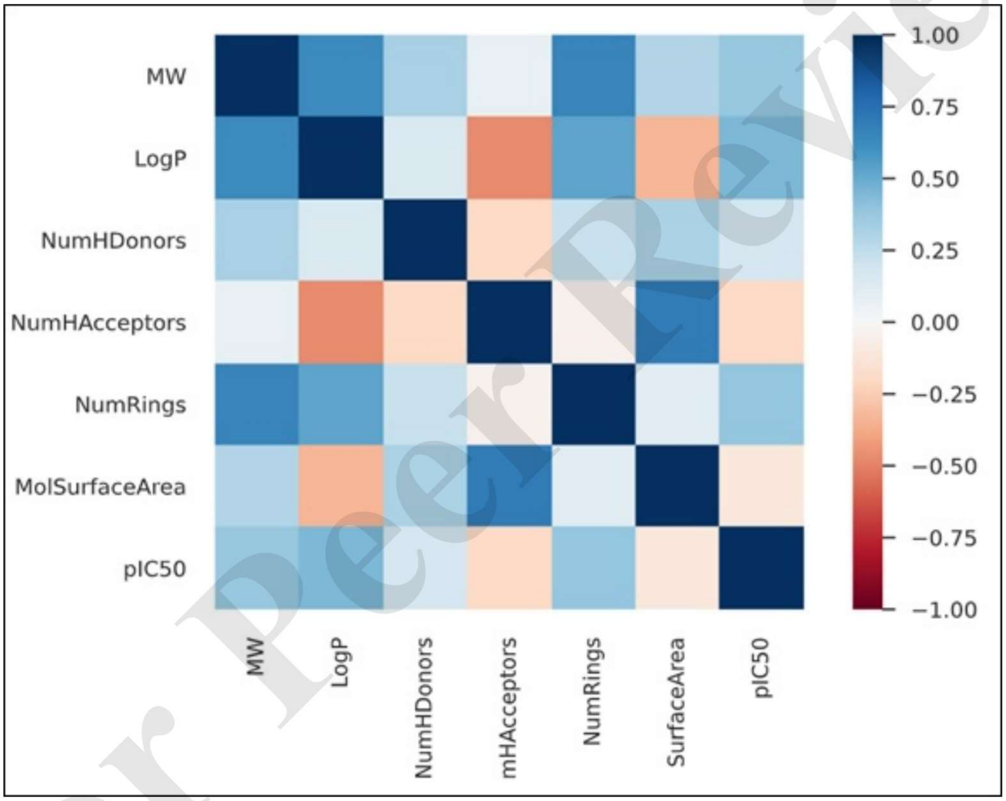
Table 2. Model Algorithm Comparison Analysis

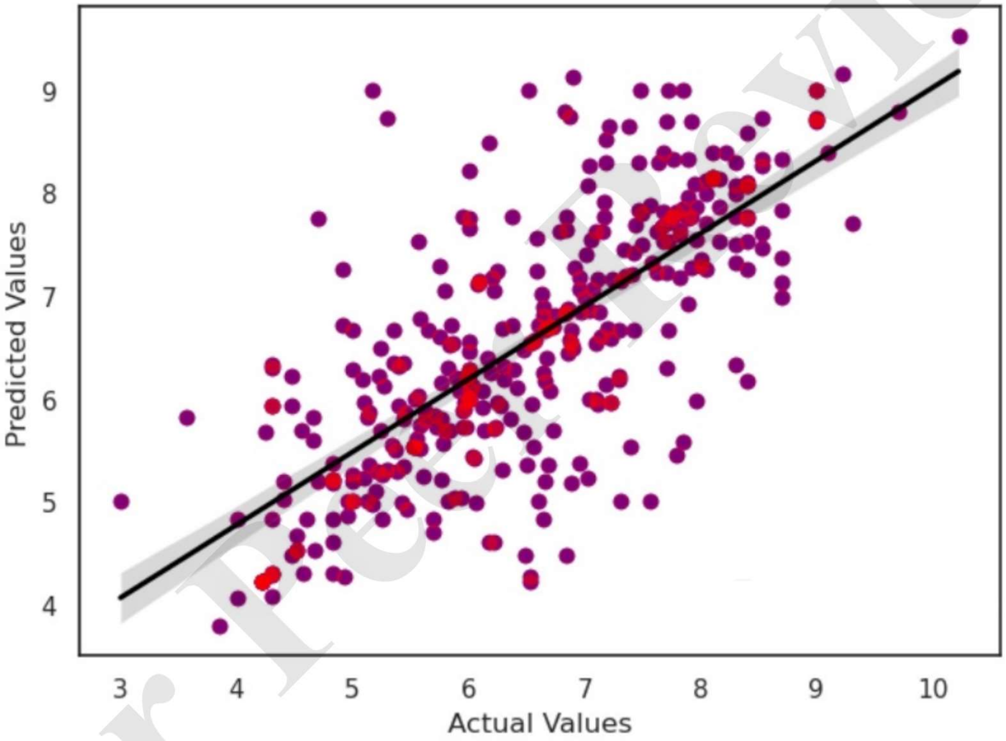
Acronym	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.5823	0.5782	0.7616	0.5999	0.1102	0.0950	0.598
rf	Random Forest Regressor	0.6136	0.6914	0.8292	0.5637	0.1121	0.0989	2.3213
gbr	Gradient Boosting Regressor	0.6847	0.7591	0.8674	0.524	0.1182	0.1107	1.148
xgboost	Extreme Gradient Boosting	0.6402	0.756	0.8676	0.5205	0.117	0.1024	3.4
br	Bayesian Ridge	0.7009	0.7986	0.8897	0.4974	0.1219	0.1136	0.816
knn	K Neighbors Regressor	0.6953	0.8368	0.9109	0.472	0.1236	0.1127	0.3947
ridge	Ridge Regression	0.7306	0.8848	0.9381	0.4399	0.13	0.1191	0.3673
omp	Orthogonal Matching Pursuit	0.7414	0.8926	0.9406	0.4392	0.1297	0.1203	0.4693
huber	Huber Regressor	0.734	0.9216	0.9569	0.4146	0.1336	0.12	0.7807
et	Extra Trees Regressor	0.6765	0.9238	0.959	0.4141	0.1295	0.1081	2.842
ada	AdaBoost Regressor	0.8179	0.9661	0.9796	0.3922	0.1343	0.1335	0.7067
dt	Decision Tree Regressor	0.6963	0.9898	0.9925	0.371	0.134	0.111	0.518
en	Elastic Net	1.0394	1.5575	1.2443	0.0272	0.17	0.1704	0.5067
lasso	Lasso Regression	1.0617	1.6252	1.2712	-0.0154	0.1735	0.174	0.328
llar	Lasso Least Angle Regression	1.0617	1.6252	1.2712	-0.0154	0.1735	0.174	0.5267
dummy	Dummy Regressor	1.0617	1.6252	1.2712	-0.0154	0.1735	0.174	0.5013
par	Passive Aggressive Regressor	1.0268	1.7591	1.3136	-0.1196	0.1842	0.1649	0.3627
lr	Linear Regression	1018	6.52	9337	-3.419	1.035	1492	4.5613
lar	Least Angle Regression	5.227	1.374	9.573	-7.295	14.07	9.528	0.5753

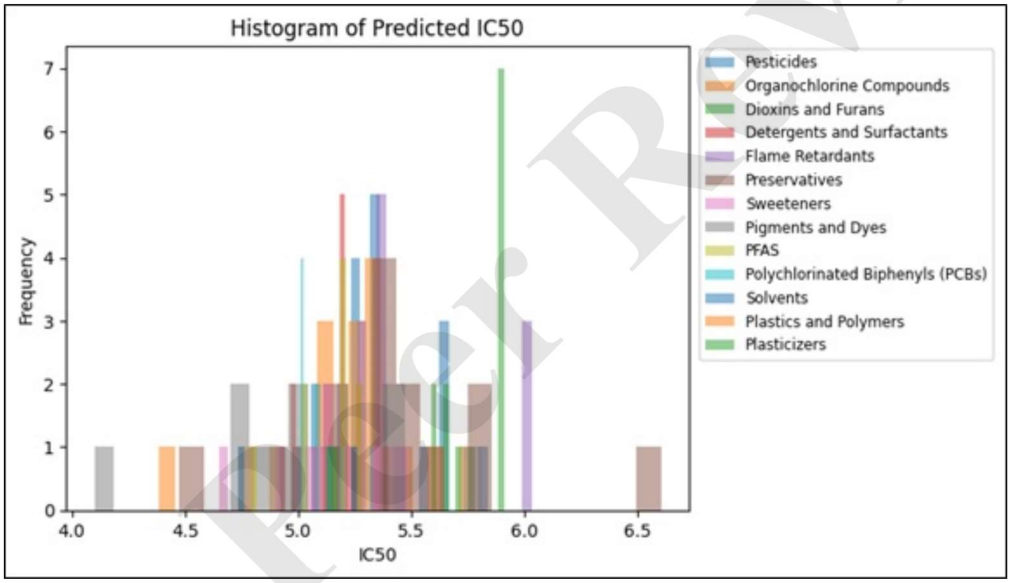
Table 3. Performance metrics for the LightGBM model, detailing training and testing scores. Metrics include R², MSE, MAE, RMSE, RMSLE, and MAPE, which collectively demonstrate the model's robustness and accuracy in predicting IC50 values for PPARγ inhibition

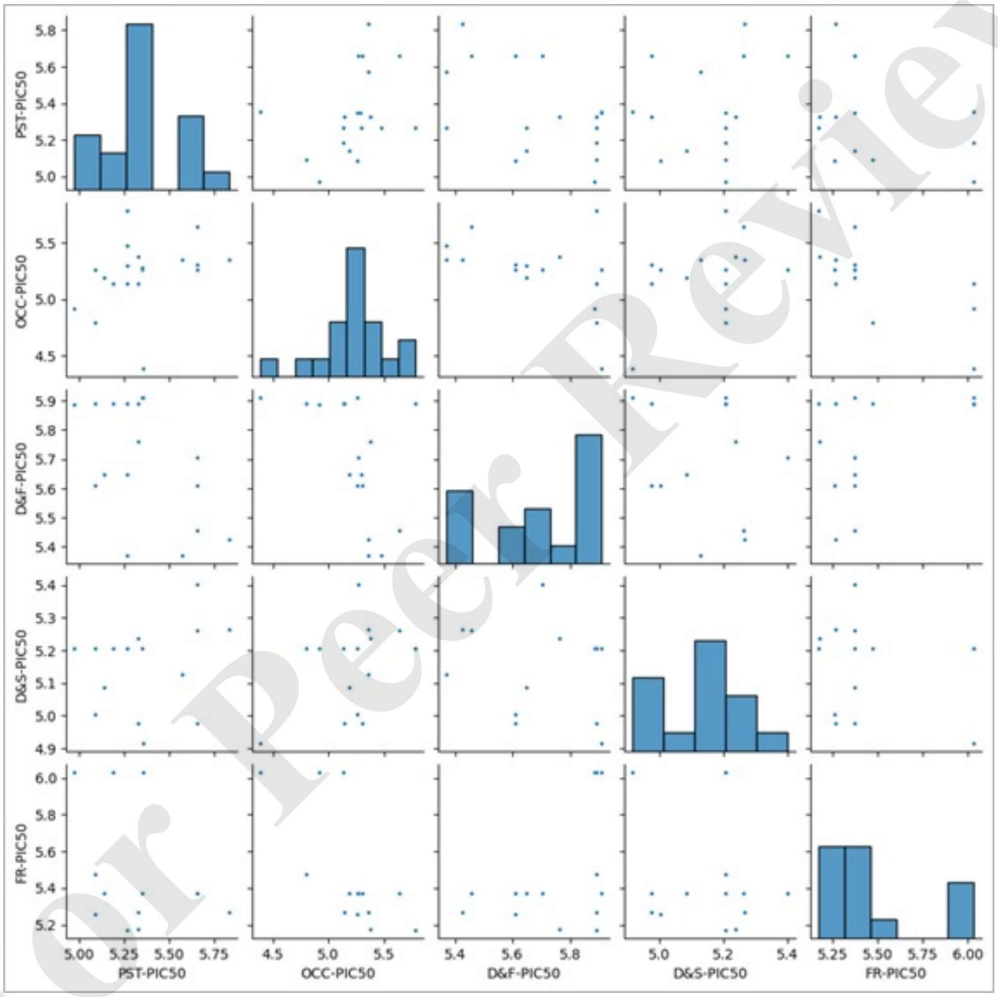
Metric	Training Score	Testing Score
R^2	0.82	0.60
MSE	0.29	0.58
MAE	0.38	0.58
RMSE	0.54	0.76
RMSLE	0.07	0.11
MAPE	0.06	0.09

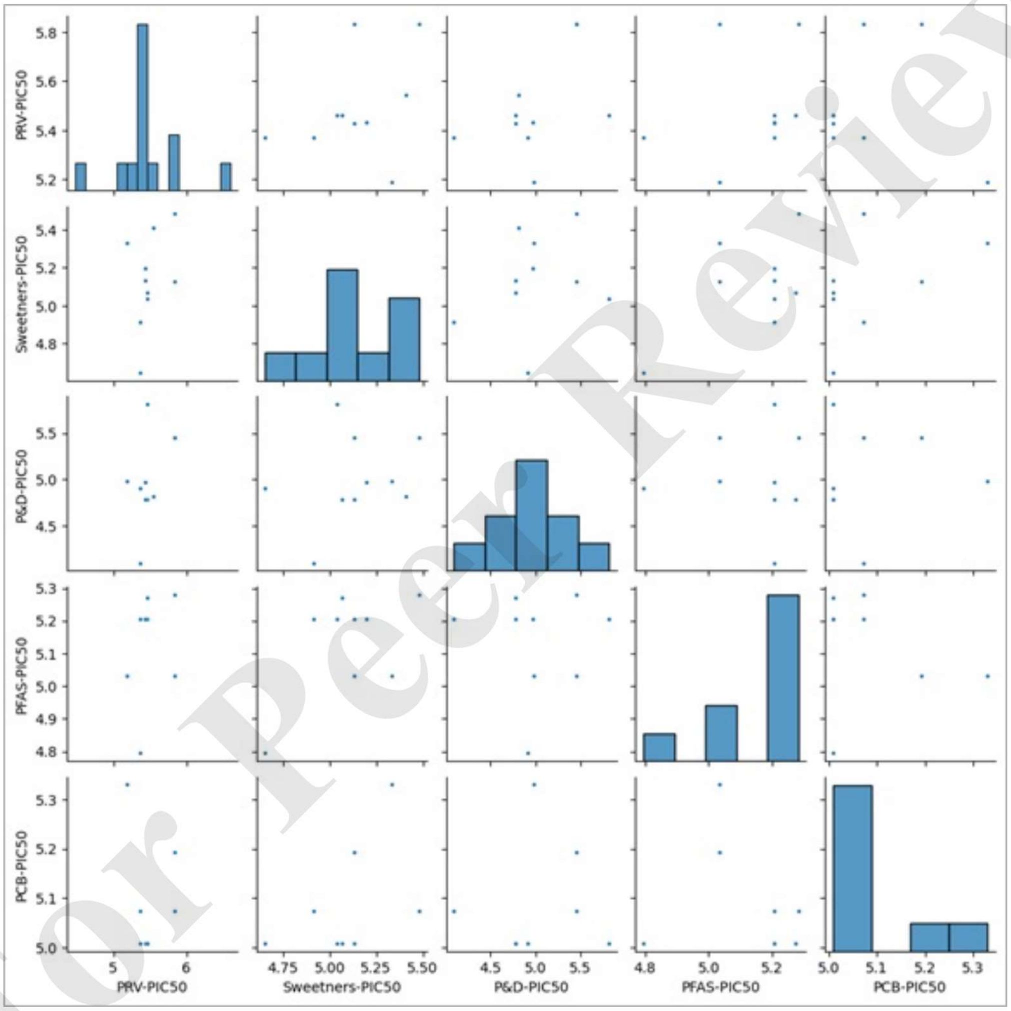

Validation Curve for LGBMRegressor
Training Score
Cross Validation Score
score
0.8
0.7
0.6
0.5
0.4
0.3
2
4
6
8
10
max_depth

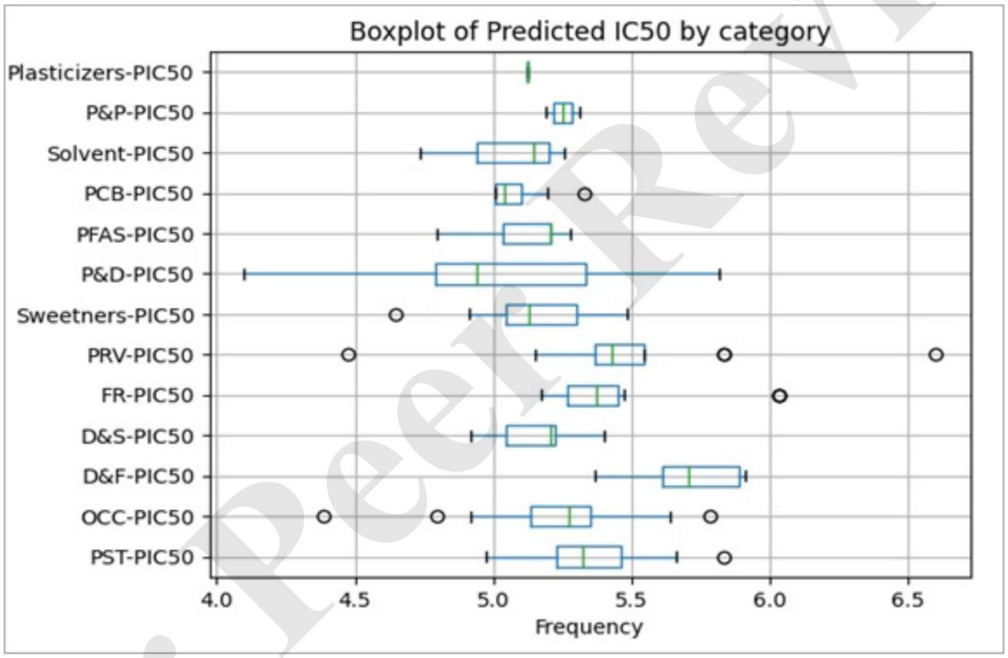

MW
LogP
NumHDonors
NumHAcceptors
NumRings
MolSurfaceArea
pIC50
MW
LogP
NumHDonors
mHAcceptors
NumRings
SurfaceArea
pIC50
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Predicted Values
Actual Values
4
5
6
7
8
9
3
4
5
6
7
8
9
10


Histogram of Predicted IC50
Frequency
0
1
2
3
4
5
6
7
4.0
4.5
5.0
5.5
6.0
6.5
IC50
Pesticides
Organochlorine Compounds
Dioxins and Furans
Detergents and Surfactants
Flame Retardants
Preservatives
Sweeteners
Pigments and Dyes
PFAS
Polychlorinated Biphenyls (PCBs)
Solvents
Plastics and Polymers
Plasticizers


PST-PIC50
OCC-PIC50
D&F-PIC50
D&S-PIC50
FR-PIC50
5.8
5.6
5.4
5.2
5.0
5.5
5.0
4.5
5.9
5.8
5.7
5.6
5.5
5.4
5.4
5.3
5.2
5.1
5.0
4.9
6.0
5.8
5.6
5.4
5.2
5.00 5.25 5.50 5.75
4.5 5.0 5.5
5.4 5.6 5.8
5.0 5.2 5.4
5.25 5.50 5.75 6.00
PST-PIC50
OCC-PIC50
D&F-PIC50
D&S-PIC50
FR-PIC50


PRV-PIC50
5.2
5.4
5.6
5.8
Sweetners-PIC50
4.8
5.0
5.2
5.4
P&D-PIC50
4.5
5.0
5.5
PFAS-PIC50
4.8
4.9
5.0
5.1
5.2
5.3
PCB-PIC50
5.0
5.1
5.2
5.3
5
6
PRV-PIC50
4.75
5.00
5.25
5.50
Sweetners-PIC50
4.5
5.0
5.5
P&D-PIC50
4.8
5.0
5.2
PFAS-PIC50
5.0
5.1
5.2
5.3
PCB-PIC50


Boxplot of Predicted IC50 by category
Plasticizers-PIC50
P&P-PIC50
Solvent-PIC50
PCB-PIC50
PFAS-PIC50
P&D-PIC50
Sweetners-PIC50
PRV-PIC50
FR-PIC50
D&S-PIC50
D&F-PIC50
OCC-PIC50
PST-PIC50
4.0
4.5
5.0
5.5
6.0
6.5
Frequency