chemistry

MDPI

Article

Application of Machine Learning Models for Predicting pIC50 Values of Plasticizers Against Cytochrome P450 Aromatase

Itumeleng Lucky Mongadi [1], Nomasonto Rapulenyane [1,*], Walter Bonke Mahlangu [2,*] and Jean-Nazaire Oyourou [2,*]

[1] Department of Chemistry and Chemical Technology, Sefako Makgatho Health Sciences University, Medunsa 0204, South Africa
[2] Department of Applied Sciences, Eduvos, 44 Alsatian Rd, Glen Austin AH, Midrand 1685, South Africa
* Correspondence: nomasonto.rapulenyane@smu.ac.za (N.R.); bonke.mahlangu@eduvos.com (W.B.M.); jean.oyourou@eduvos.com (J.-N.O.)

Abstract

This study investigated the application of six machine learning regression algorithms such as Random Forest, CatBoost, K-Nearest Neighbours, XGBoost, LightGBM, and Gradient Boosting, paired with Molecular ACCess System (MACCS) key fingerprints for the quantitative prediction of aromatase (CYP19A1) inhibitory potency, expressed as pIC_{50}. A dataset of 187 compounds was assembled from the ChEMBL database (version 33, Target ID: CHEMBL1978) following by systematic data curation workflow encompassing duplicate removal, pIC_{50} transformation, and activity-based filtering. Model performance was rigorously evaluated using an 80/20 stratified train/test split, 5-fold cross-validation, and Y-randomisation testing to ensure unbiased assessment of predictive generalisation. Feature selection via CatBoost permutation importance on the held-out test set identified the top 20 predictive MACCS keys from an initial 166-bit space, substantially reducing dimensionality and improving generalisation across all models. Among the algorithms evaluated, CatBoost trained on the top 20 features achieved the strongest test-set performance (R^2 = 0.693, RMSE = 0.794, MAE = 0.659) with the most stable cross-validation R^2 (0.062 $\pm$ 0.304), outperforming all other algorithms. Y-randomisation testing returned an empirical *p*-value of <0.01, confirming that model performance reflects genuine structure–activity relationships rather than statistical chance. Permutation importance and SHAP analysis identified nitrogen-containing heterocyclic fragments (MACCS_41, MACCS_145) and halide-bearing substructures (MACCS_109) as the primary structural determinants of aromatase inhibitory potency, consistent with established CYP19A1 pharmacophoric requirements. Application of the model to ten representative plasticizers demonstrated that the refined applicability domain (h* = 0.423) accommodated eight of the ten compounds, enabling reliable potency predictions across phthalate esters and bisphenol analogues. These findings establish a transparent and reproducible QSAR framework for first-tier endocrine disruption risk screening of plasticizers and highlight the importance of permutation-based feature selection and applicability domain assessment in QSAR model development.

Keywords: cytochrome P450 aromatase; quantitative structure–activity relationship (QSAR); machine learning regression algorithms; molecular ACCess system keys (MACCS); estrogen biosynthesis inhibition

Check for updates

Academic Editors: Jun Li, Tong Zhu and Valentine P. Ananikov

Received: 23 March 2026
Revised: 10 April 2026
Accepted: 22 April 2026
Published: 20 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Plastics are widely used in modern society due to their versatility and cost-effectiveness. To enhance their elasticity and flexibility, manufacturers incorporate plasticizers into various products, including toys, medical equipment, and food packaging [1]. However, some plasticizers, such as bisphenols, have raised significant health concerns due to their potential links to cancer and reproductive complications [2]. Polycarbonate plastic, a widely used material, was first synthesized in the 1950s using bisphenol A (BPA). This innovation allowed manufacturers to create a range of products, including food containers, packaging materials, feeding bottles, dinnerware, and technological devices [3]. BPA also serves as a critical intermediate in the production of epoxy resins used for coatings in metal cans and bottle caps, adhesives, medical equipment, and thermally printed papers [4]. Despite its widespread utility, BPA has been shown to leach into food and beverages from containers made with this compound. This leaching has raised concerns about potential health risks associated with BPA exposure [5].

In response to the growing public awareness of BPA's health risks, manufacturers began substituting BPA with alternative chemicals such as bisphenol S (BPS) and bisphenol F (BPF). However, studies have detected traces of BPA, BPS, and BPF in human urine samples, indicating that exposure to these chemicals remains an ongoing issue [6]. This suggests that replacing BPA with structurally similar compounds may not fully address the underlying health risks. In addition to bisphenols, phthalates represent another group of plasticizers with significant environmental and health impacts. Common phthalates include di-n-butyl phthalates (DnBP), benzyl butyl phthalate (BBP), and diethylhexyl phthalate (DEHP) [7,8]. These chemicals are widely used in plastic products and can contaminate water sources through various pathways, including direct discharge from industrial processes, effluents from wastewater treatment plants, landfill leachates, and surface runoff from rainwater. Even treated drinking water may contain trace amounts of these chemicals due to the use of plastic pipes and storage containers in water distribution systems [9].

In addition to their environmental implications, some plasticizers have been implicated in disrupting critical biological processes such as steroidogenesis, an area where cytochrome P450 aromatase plays a central role [10]. Cytochrome P450 aromatase (CYP19A1) is a crucial enzyme in the biosynthesis of oestrogens from androgen precursors. It catalyzes the conversions of androstenedione into oestrogen and testosterone into oestradiol through a series of hydroxylations, followed by demethylation and aromatization of the A-ring [10,11]. Aromatase is localized in the endoplasmic reticulum and requires NADPH-cytochrome P450 reductase for electron transfer during catalysis [10].

The enzyme is expressed in various tissues, including the ovaries, placenta, adipose tissue, brain, and blood vessels. In the brain, aromatase plays a role in neuroprotection, neurogenesis, and modulation of emotional states and cognitive functions [12]. Its expression in neurons and reactive expression in astrocytes under pathological conditions highlights its importance in neural health [12]. Aromatase exhibits a highly substrate-selective nature, optimized for catalysis on androgens. Comparative sequences and structural analysis have identified conserved amino acids essential for its function, such as those involved in the proton relay network during catalysis [11,13]. These conserved residues have remained unchanged throughout evolution, underscoring their critical roles in enzymatic activity.

The potency of an inhibitor is commonly assessed by its IC_{50} value, which is usually in molar (M), micromolar (μM), or nanomolar (nM). It represents the concentration required to inhibit 50% of the target activity. A lower IC_{50} value indicates a more potent inhibitor, while a higher IC_{50} value indicates a less potent inhibitor [14]. Aromatase inhibitors are classified into two main types based on their chemical structure and mechanism of action.

1.1. Non-Steroidal Aromatase Inhibitors

Non-steroidal inhibitors include compounds like letrozole and anastrozole, which are third-generation compounds that reversibly bind to the enzyme's active site [15]. These inhibitors act by attaching to the heme group of the aromatase via their triazole moiety, preventing the conversion of androgens into oestrogens. Letrozole is a highly potent inhibitor, with IC_{50} values ranging from approximately 50–100 nM in various vitro models of oestrogen receptor-positive breast cancer cells [16]. Anastrozole, while also effective, exhibits slightly lower potency compared to letrozole, with IC_{50} values typically exceeding 100 nM under similar conditions [15].

1.2. Steroidal Aromatase Inhibitors

Steroidal aromatase inhibitors, such as exemestane, are classified as irreversible due to their unique mechanism of action. These compounds mimic the natural androgen substrates of aromatase and bind covalently to the enzyme's active site, leading to its permanent inactivation. This irreversible binding results in sustained suppression of oestrogen synthesis, making exemestane particularly effective in managing oestrogen-dependent conditions like hormone-receptor-positive breast cancer [15].

With the aid of quantitative structure-activity relationship (QSAR), the prediction of biological activities for investigated compounds can be achieved using machine learning and artificial intelligence [17]. QSAR models can also be used in toxicity assessment, such as the potential to cause cancer and other health complications. This can help in prioritizing chemicals for testing and assessing the risk of exposure to chemicals. This study develops and rigorously validates a QSAR framework employing six machine learning regression algorithms paired with MACCS key fingerprints, permutation-based feature selection, 5-fold cross-validation, Y-randomisation, and applicability domain assessment to predict the pIC50 values of plasticizers against cytochrome P450 aromatase (CYP19A1), enabling reliable and mechanistically interpretable first-tier endocrine disruption risk screening.

statement of significance:

This study employed machine learning models to predict the pIC_{50} values of plasticizers against cytochrome P450 aromatase (CYP19A1), a central enzyme in oestrogen biosynthesis whose inhibition by environmental chemicals poses significant endocrine disruption risk. Plasticizers such as bisphenols and phthalates are ubiquitous in consumer products and have been detected in human biological matrices, yet their aromatase inhibitory potential remains poorly characterised at scale. Conventional experimental screening of plasticizers for aromatase inhibitory activity is resource-intensive, time-consuming, and constrained by the sheer number of compounds requiring evaluation. This study addresses this challenge by developing and rigorously validating a machine learning QSAR framework, employing CatBoost with permutation-based feature selection from MACCS fingerprints, 5-fold cross-validation, Y-randomisation, SHAP interpretability, and applicability domain assessment, capable of providing reliable and mechanistically interpretable pIC50 predictions for compounds within its defined chemical space. By applying comprehensive validation protocols, this work establishes a transparent and reproducible QSAR framework that can serve as a first-tier screening tool for prioritising plasticizers warranting further experimental investigation. Feature importance and SHAP analyses provide mechanistic insight into the structural features associated with aromatase inhibition, contributing to a molecular-level understanding of endocrine disruption that can inform the rational design of safer plasticizer alternatives. The applicability domain findings furthermore demonstrate that the dimensionality-reduced model accommodates the majority of evaluated plasticizer scaffolds, substantially expanding its practical utility for chemical risk assessment. Crucially, the applicability domain framework provides the

scientific transparency required for regulatory acceptance under OECD QSAR validation principles. The study creates urgency for expanding training datasets with experimentally characterised plasticizer-specific bioactivity data to extend the model's predictive reach and highlights the need for evidence-based regulation of plasticizers to limit hormone-mediated toxicological effects on human health and environmental systems.

2. Methodology

2.1. Dataset

The homo sapiens aromatase inhibitor dataset (ChEMBL Target ID: CHEMBL1978) was retrieved from the ChEMBL database (version 33) [18]. Records were filtered using standard type = IC_{50}, standard units = nM, and standard relation = '=', returning 330 records. IC_{50} values were converted to pIC_{50} using $pIC_{50} = -\log_{10}(IC_{50} \times 10^{-9})$ to normalise the distribution and facilitate structure–activity relationship modelling.

2.2. Data Curation Workflow

A sequential data curation workflow was applied to ensure data quality and reproducibility (Table 1). Records with missing IC_{50} values or null canonical SMILES were removed ($n = 8$), yielding 322 records. Duplicate structures were identified by comparing canonical SMILES strings, retaining one record per unique structure (249 unique compounds). Compounds with IC_{50} values between 1000 and 10,000 nM were classified as 'intermediate' and excluded ($n = 62$), as their ambiguous activity introduces noise at class boundaries. It should be noted that these activity labels (active: $IC_{50} < 1000$ nM; inactive: $IC_{50} > 10{,}000$ nM) were used exclusively as a data curation filter and were not employed as target variables in regression model training. The final dataset comprised 187 compounds (139 active, 48 inactive).

Table 1. Data curation workflow—Aromatase inhibitor dataset (ChEMBL v33, ID: CHEMBL1978).

Step	Description	Criterion/Action	Compounds Remaining	Compounds Removed	Ref.
1	Raw ChEMBL records (v33, CHEMBL1978)	Retrieve: IC_{50}, nM, relation '='	330	—	[18]
2	Remove missing values	Drop null IC_{50} or null canonical SMILES	322	8	
3	Remove duplicate structures	One record per unique canonical SMILES	249	73	
4	Exclude intermediate-activity compounds	Remove $1000 < IC_{50} \leq 10{,}000$ nM	187	62	
5	pIC_{50} transformation	$pIC_{50} = -\log_{10}(IC_{50} \times 10^{-9})$	187	0	
6	Generate MACCS fingerprints	RDKit MACCSkeys (166 bits)	187	0	[19]
	Final dataset for modelling		**187**	—	

2.3. Molecular Descriptors

Molecular ACCess System (MACCS) keys were computed for each compound using RDKit (v2023.09.1) [19] based on canonical SMILES. MACCS keys comprise a fixed 166-bit binary fingerprint where each bit encodes the presence or absence of a specific structural fragment. The resulting descriptor matrix (187×166) served as the input feature set for all models.

2.4. Train/Test Split

The curated dataset (n = 187) was partitioned into a training set (80%, n = 149) and a held-out test set (20%, n = 38) using stratified random sampling based on tertile binning of pIC_{50} values, preserving the pIC_{50} distribution across both sets (training mean: 6.06 ± 1.90; test mean: 6.13 ± 1.43). The test set was held out exclusively for final model evaluation and was not exposed during training or cross-validation.

2.5. Model Training Model Training and Hyperparameters

Six regression algorithms were trained: Random Forest (RF), CatBoost (CB), K-Nearest Neighbours (KNN), XGBoost (XGB), LightGBM (LGBM), and Gradient Boosting (GB). Table 2 lists the hyperparameters used for each model.

Table 2. Final hyperparameters for each regression model.

Model	Key Hyperparameters	Default/Tuned
Random Forest	n_estimators = 100, random_state = 42, n_jobs = −1	Default
CatBoost	iterations = 300, learning_rate = 0.05, depth = 6, random_seed = 42	Default
KNN	n_neighbors = 5, metric = 'minkowski'	Default
XGBoost	n_estimators = 200, learning_rate = 0.05, max_depth = 4, random_state = 42	Default
LightGBM	n_estimators = 200, learning_rate = 0.05, num_leaves = 31, random_state = 42	Default
Gradient Boosting	n_estimators = 200, learning_rate = 0.05, max_depth = 4, random_state = 42	Default

2.6. Model Validation

Model performance was assessed using three complementary validation approaches, with three metrics applied consistently throughout.

Evaluation Metrics. The coefficient of determination (R^2) measures the proportion of variance in pIC_{50} values that the model successfully explains. An R^2 of 1.0 represents perfect prediction, 0.0 indicates the model performs no better than simply predicting the mean pIC_{50} for every compound, and negative values reveal that the model performs worse than the mean, a direct indicator of overfitting or failed generalisation to unseen data. The root mean squared error (RMSE) quantifies the average magnitude of prediction errors in pIC_{50} units, penalising large individual errors disproportionately due to the squaring operation, making it particularly sensitive to outlier predictions. The mean absolute error (MAE) represents the average absolute difference between predicted and experimental pIC_{50} values, treating all errors equally regardless of their size, and therefore provides a more robust and directly interpretable measure of typical everyday prediction accuracy. Used together, R^2 describes overall model fit quality, RMSE flags occasional large prediction failures, and MAE characterises the model's routine predictive precision.

(i) *Train/test evaluation:* All models were evaluated on both the training set and the held-out test set using R^2, RMSE, and MAE. A large discrepancy between training and test metrics is a direct indicator of overfitting, the model has memorised patterns in the training data rather than learning transferable structure–activity relationships that generalise to new compounds.

(ii) *Five-fold cross-validation:* The training set was partitioned into five folds (KFold, shuffle = True, random_state = 42) to obtain mean ± SD estimates of R^2, RMSE, and MAE across folds. The mean CV R^2 reflects average generalisation performance across different subsets of the training data. Critically, the standard deviation of CV R^2 serves as a stability indicator, a large SD signals that model performance depends heavily on which specific compounds fall in each fold, exposing sensitivity to training

set composition rather than robust learning of underlying chemical patterns. This is particularly informative for small datasets where individual compounds can have an outsized influence on model behaviour.

(iii) Y-randomisation: pIC_{50} values in the training set were randomly permuted 100 times and the CatBoost model was independently retrained and evaluated on the test set for each permutation. If the real model's performance substantially exceeds that of all permuted models, it confirms that predictive ability arises from genuine structure–activity relationships encoded in the MACCS descriptors rather than from statistical artefacts, dataset size, or chance correlations in the data partitioning.

2.7. Applicability Domain Assessment

The applicability domain (AD) defines the boundaries of chemical space within which a QSAR model is considered capable of making reliable predictions. A model applied to compounds that are structurally dissimilar from its training set is operating by extrapolation, and its predictions carry no statistical or chemical guarantee of accuracy. In the context of this study, where the primary objective is to assess the aromatase inhibitory risk of real-world plasticizers, the AD is of critical scientific significance: it determines which plasticizer predictions can be defended as reliable estimates and which must be treated as exploratory extrapolations requiring experimental verification. AD assessment is also a mandatory component of QSAR model reporting under OECD Principle 3, which governs the validation and regulatory acceptance of QSAR models in chemical risk assessment [20].

The AD was assessed using leverage-based analysis [20]. The leverage of each compound was calculated as $h_i = x_i^T(X^TX)^{-1}x_i$, where x_i is the compound's MACCS fingerprint vector for the top 20 selected features and X is the training set descriptor matrix. Leverage quantifies how structurally distant a compound is from the centroid of the training set in descriptor space: a compound with high leverage occupies a region poorly represented by the training data, meaning the model has little chemical basis for its prediction and that prediction is inherently unreliable. The warning leverage threshold was recalculated for the top-20 feature model as $h^* = 3(k + 1)/n = 3(20 + 1)/149 = 0.423$, where $k = 20$ is the number of features and $n = 149$ is the number of training compounds. This threshold is substantially tighter than the full-feature baseline value of 3.383, reflecting the more focused and geometrically compact descriptor space defined by 20 MACCS keys compared to the full 166-bit space. Compounds with standardised residuals outside $\pm 3\sigma$ were additionally flagged as response outliers, indicating that even if a compound falls within the leverage boundary, an unusually large prediction error suggests the model cannot adequately describe its activity. A Williams plot, a scatter plot of leverage against standardised residuals, was constructed to simultaneously visualise both the structural and response-based boundaries of the AD, enabling clear identification of compounds that are reliable predictions, structural outliers, response outliers, or both.

2.8. Feature Importance Analysis

Feature importance was assessed using two complementary methods applied to the best-performing model, CatBoost. First, permutation importance was computed on the held-out test set: each of the 166 MACCS bits was independently shuffled 30 times, and the resulting decrease in test-set R^2 was recorded as the importance score. This approach directly measures each feature's contribution to predictive generalisation on unseen data, avoiding the bias inherent in training-set-based methods such as mean decrease in impurity (MDI). The top 20 MACCS keys ranked by permutation importance were selected as the final feature subset for model retraining. Second, SHAP (SHapley Additive exPlanations) values were computed for the CatBoost model retrained on the top 20 features

using the shap. TreeExplainer interface. SHAP values decompose each individual prediction into additive feature contributions, revealing not only which features are important globally but the direction and magnitude of their effect on each compound's predicted pIC_{50}. Structural interpretations of the identified MACCS keys were obtained via rdkit.Chem.MACCSkeys.smartsPatterns.

3. Results and Discussion

3.1. Model Performance: Training, Cross-Validation, and Test Set

Table 3 presents the comprehensive performance metrics for all six regression models trained on the top 20 MACCS keys identified by CatBoost permutation importance, evaluated across the training set, 5-fold cross-validation, and the held-out test set. Figure 1 shows predicted versus experimental pIC_{50} scatter plots for all models on the test set. Figure 2 presents a comparative bar chart of CV R^2 and test set R^2 across the baseline (all 166 bits) and top feature subsets (top 10, 20, and 30), demonstrating that training on the top 20 features consistently improved both predictive performance and cross-validation stability across all models.

Table 3. Comprehensive model performance metrics trained on top 20 MACCS keys (187 compounds; 80/20 stratified split; 5-fold CV). ★ Indicates best test-set R^2.

Model	Train R^2	Train RMSE	Train MAE	CV R^2 (Mean ± SD)	CV RMSE (Mean ± SD)	Test R^2	Test RMSE	Test MAE
Random Forest	0.650	1.127	0.435	−0.226 ± 0.775	1.859 ± 0.529	0.551	0.959	0.746
CatBoost ★	0.732	0.986	0.367	0.062 ± 0.304	1.744 ± 0.587	0.693	0.794	0.659
KNN	0.375	1.505	0.995	0.008 ± 0.265	1.854 ± 0.547	0.456	1.057	0.835
XGBoost	0.711	1.024	0.327	−0.640 ± 1.687	2.165 ± 0.720	0.432	1.079	0.777
LightGBM	0.297	1.597	0.682	−0.081 ± 0.202	1.855 ± 0.520	0.324	1.178	1.008
Gradient Boosting	0.732	0.985	0.233	−0.660 ± 1.676	2.175 ± 0.714	0.326	1.176	0.807

Following permutation-based feature selection, all six algorithms were retrained on the top 20 MACCS keys and re-evaluated across training, cross-validation, and held-out test sets (Table 3). The dimensionality reduction consistently improved both predictive generalisation and cross-validation stability relative to the full 166-bit descriptor space. Across all models, a divergence between training-set performance and test-set generalisation remained, a pattern characteristic of QSAR regression on datasets of moderate size and chemical diversity [21]. Random Forest achieved a training R^2 of 0.650 and a test-set R^2 of 0.551 (RMSE = 0.959, MAE = 0.746), representing a substantially improved generalisation performance compared to the baseline full-feature model. XGBoost and Gradient Boosting continued to exhibit the most pronounced train–test discrepancies, with training R^2 values of 0.711 and 0.732 declining to 0.432 and 0.326 on the test set respectively, and their negative cross-validation R^2 values (XGBoost: -0.640 ± 1.687; Gradient Boosting: -0.660 ± 1.676) indicate high fold-to-fold variance, underscoring the sensitivity of these high-capacity learners to training subset composition at this sample size. CatBoost demonstrated the most favourable balance between training performance and test-set generalisation, achieving the highest test-set R^2 (0.693), the lowest test RMSE (0.794) and MAE (0.659), and the most stable cross-validation R^2 (0.062 ± 0.304), attributable to its ordered boosting procedure and built-in regularisation mechanisms that reduce target leakage during training [22]. KNN achieved a test-set R^2 of 0.456 (RMSE = 1.057, MAE = 0.835), a notable improvement over the full-feature baseline, suggesting that the reduced descriptor space removed noisy bits

that impaired nearest-neighbour distance calculations. LightGBM achieved a test-set R^2 of 0.324, remaining the lowest-performing algorithm, consistent with its known sensitivity to dataset size and the relatively limited training set available in this study. Collectively, the top-20 feature models demonstrate that targeted dimensionality reduction translates directly into improved and more consistent predictive generalisation across all algorithms, with CatBoost establishing itself as the most reliable framework for pIC_{50} prediction within this chemical space.

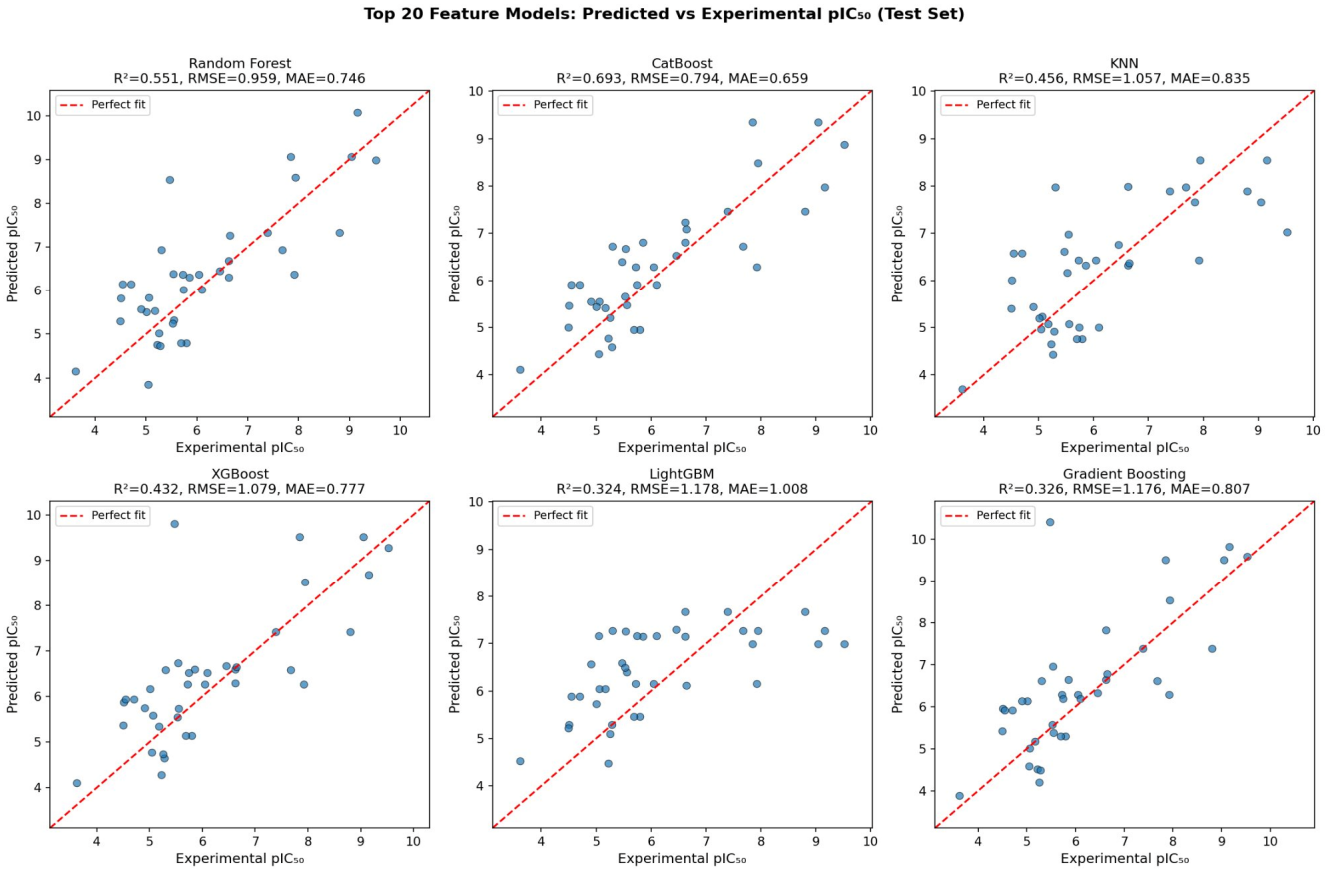


Figure 1. Predicted versus experimental pIC_{50} values for all six regression models evaluated on the held-out test set (n = 38, 20% of dataset). The dashed line represents perfect prediction (slope = 1, intercept = 0). R^2, RMSE, and MAE values are calculated on the test set only.

3.2. Y-Randomisation Test

Y-randomisation was employed to assess whether the predictive performance of the CatBoost model trained on the top 20 MACCS keys reflects genuine structural information encoded in the molecular fingerprints, or whether it could arise from fortuitous correlations within the dataset. This test deliberately destroys the relationship between chemical structure and biological activity by randomly shuffling the pIC_{50} labels: if models trained on scrambled labels perform comparably to the real model, it reveals that the real model was not learning chemistry but exploiting statistical noise. In this procedure, the pIC_{50} values of the training set were randomly permuted 100 times, and a CatBoost model was independently retrained on the top 20 MACCS features and evaluated on the held-out test set for each permutation. As shown in Figure 3, the real CatBoost model achieved a test-set R^2 of 0.693, whereas the 100 permuted models yielded a mean R^2 of -0.454 ± 0.490. None of the 100 permuted models achieved $R^2 \geq 0.693$, corresponding to an empirical *p*-value of <0.01. The large separation between the real model (R^2 = 0.693) and the permuted distribution (mean R^2 = -0.454) confirms that the CatBoost model's predictive ability

is rooted in genuine structure–activity relationships encoded in the selected MACCS descriptor space, and is not an artefact of dataset size, random partitioning, or overfitting to a particular train/test split.

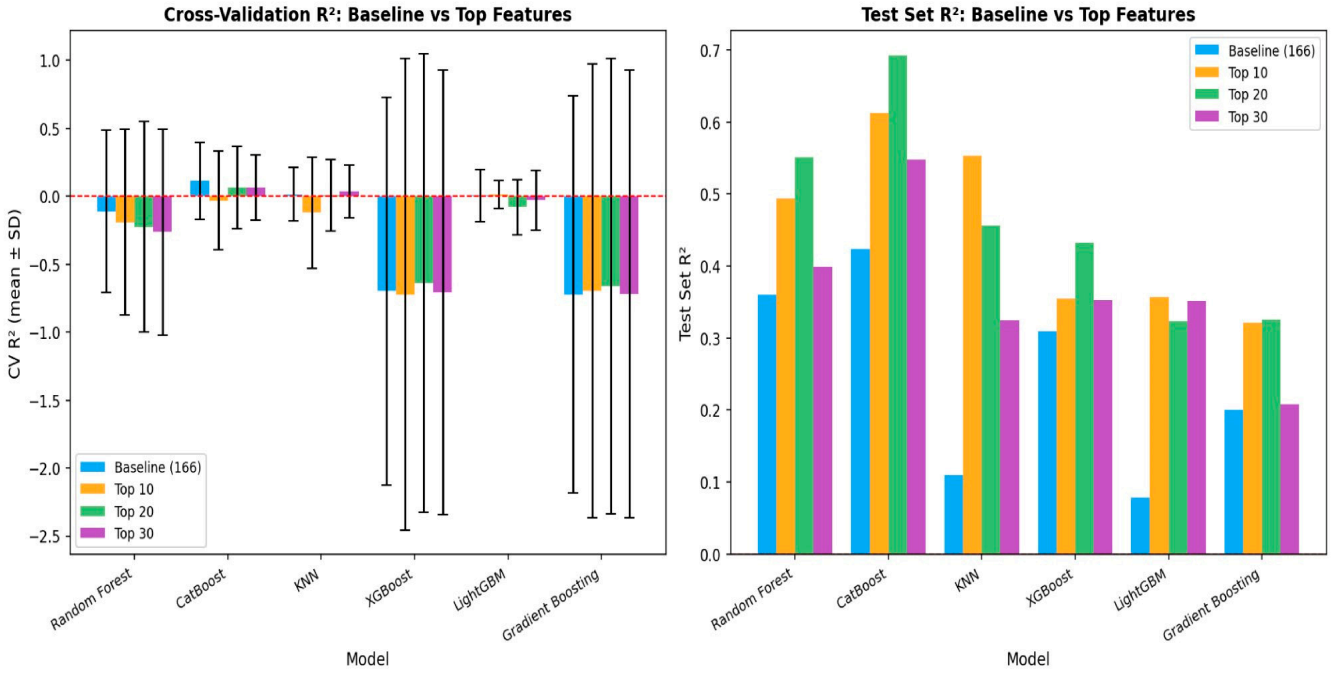


Figure 2. Comparative bar chart of R^2 values for all six models across training set (blue), 5-fold cross-validation (orange, error bars = SD), and test set (green) evaluation. The large discrepancy between training and test R^2 values illustrates the train–test generalisation gap across all models.

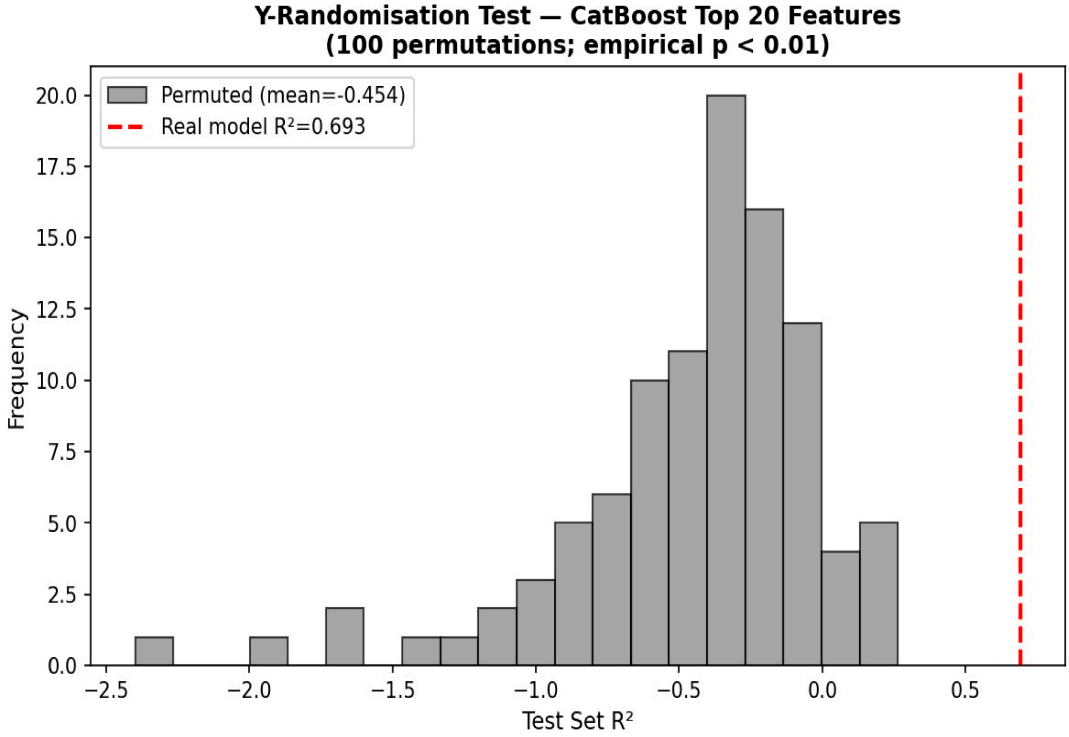


Figure 3. Y–randomisation test for the CatBoost model trained on the top 20 MACCS keys (100 permutations). The histogram shows the distribution of test-set R^2 values from models trained on randomly permuted pIC_{50} labels (grey bars; mean = -0.454 ± 0.490). The red dashed line indicates the real model R^2 (0.693). No permuted model achieved $R^2 \geq 0.693$ (empirical $p < 0.01$).

3.3. Applicability Domain Coverage and Structural Outliers

Assessment of the model's applicability domain via the Williams plot (Figure 4) revealed that the training set is internally consistent, with all 149 training compounds falling within the warning leverage threshold (h* = 0.423). Of the 38 held-out test compounds, 29 (76.3%) satisfied both the leverage criterion ($h \leq h^*$) and the residual criterion ($|\text{standardised residual}| \leq 3\sigma$), placing them within the model's applicability domain. The remaining 9 test compounds (23.7%) exhibited leverage values exceeding h*, indicating that they are structurally more dissimilar from the training set centroid than the model was designed to accommodate. This finding is informative in the context of QSAR model deployment: predictions for compounds outside the applicability domain carry greater uncertainty and should be interpreted with appropriate caution. The substantially improved applicability domain coverage reflects the more focused descriptor space defined by the top 20 MACCS keys and is a recognised advantage of dimensionality reduction in QSAR modelling when training data are drawn from public bioactivity repositories rather than congeneric series [20].

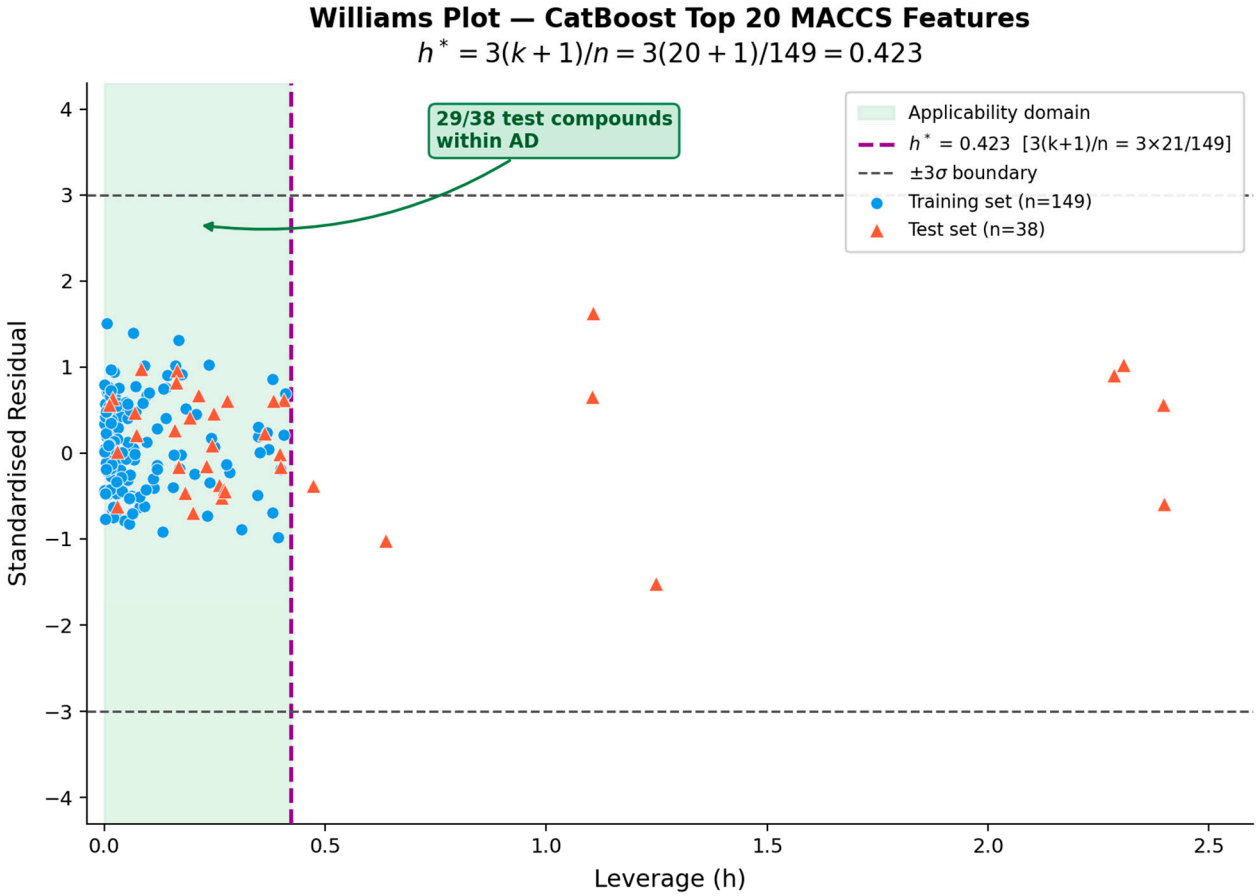


Figure 4. Williams plot showing leverage (h) versus standardised residuals for training (blue circles) and test (red triangles) compounds. The purple dashed vertical line indicates the warning leverage threshold (h* = 0.423) and grey dashed horizontal lines indicate the ±3σ residual boundaries. Compounds within both boundaries (shaded green region) are within the applicability domain.

3.4. Permutation Importance and SHAP Interpretability

Figure 5 presents the top 20 MACCS keys ranked by CatBoost permutation importance on the held-out test set, and Figure 6 presents the corresponding SHAP summary plot. MACCS_41 (mean decrease in R^2 = 0.097) and MACCS_145 (0.053) were the dominant predictors, encoding nitrogen-containing heterocyclic and pyrrole-type ring systems, re-

spectively. MACCS_122 (fused ring systems; 0.031), MACCS_109 (C-halide bonds; 0.025), and MACCS_87 (0.021) also ranked highly. SHAP analysis revealed that MACCS_41 consistently drives predicted pIC_{50} upward when present, while MACCS_109 exerts a directionally negative effect, a nuanced finding unavailable from permutation importance alone. The prominence of nitrogen heterocyclic fragments is mechanistically interpretable: azole and pyridine-type nitrogens coordinate with the haem iron of CYP19A1, a well-established pharmacophoric requirement for aromatase inhibition [14,15]. Similarly, halogenated aryl substituents are a recurring structural motif in potent non-steroidal aromatase inhibitors, including letrozole and anastrozole, where they contribute to binding affinity through hydrophobic contacts within the enzyme's active site [15,16]. The congruence between computationally derived feature importance and established pharmacophoric knowledge lends mechanistic credibility to the model and supports the biological relevance of MACCS keys as a descriptor set for this target.

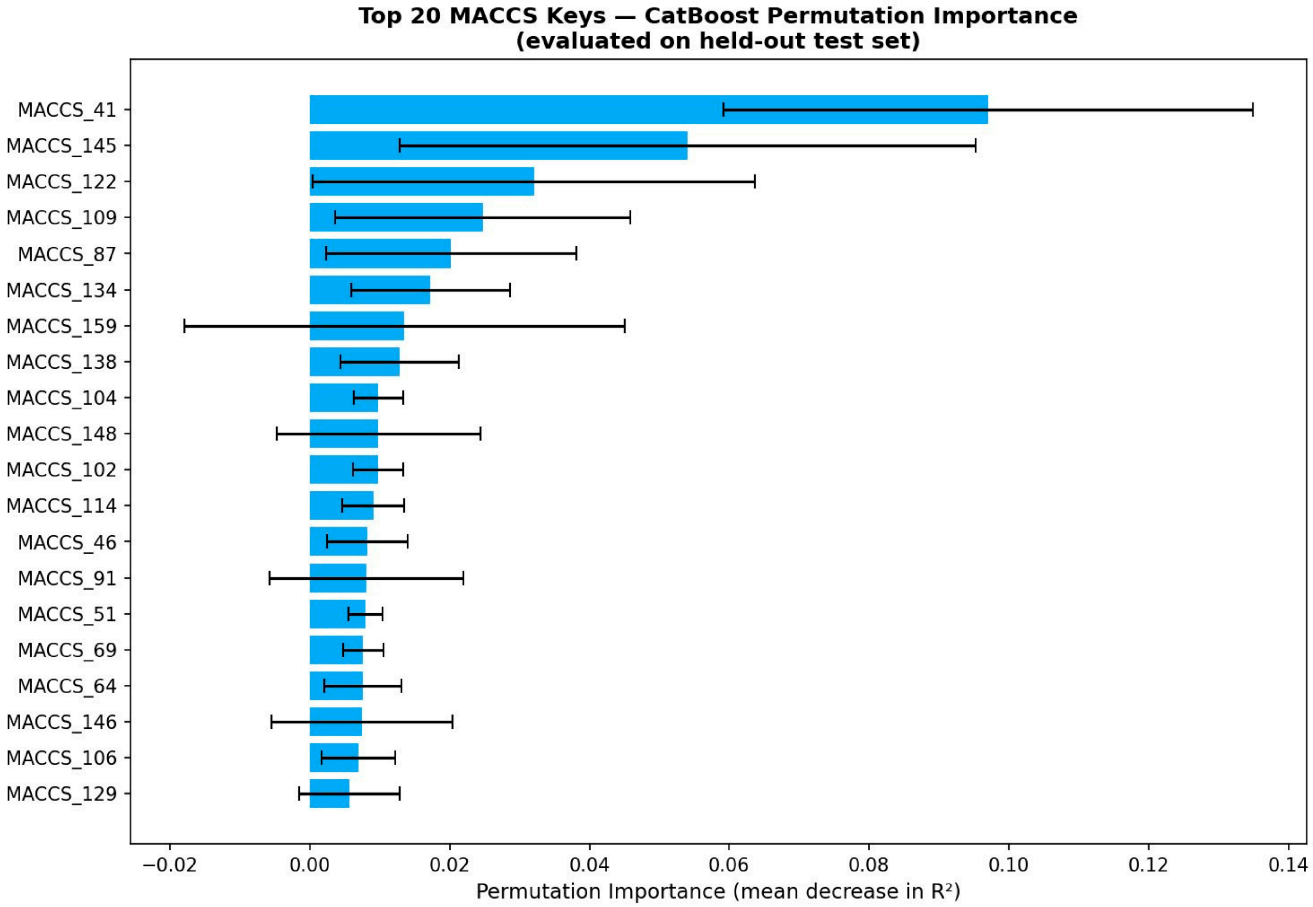


Figure 5. Top 20 MACCS keys ranked by CatBoost permutation importance (mean decrease in R^2, evaluated on the held-out test set; 30 repeats). Error bars represent ±1 standard deviation across permutation repeats. MACCS_41 and MACCS_145 are the dominant predictors, encoding nitrogen-containing heterocyclic and pyrrole-type fragments respectively.

3.5. Predicted pIC_{50} Values for Representative Plasticizers

The practical application of the developed QSAR framework was evaluated by applying the CatBoost model, trained on the top 20 permutation-selected MACCS keys, to a set of ten structurally representative plasticizers spanning phthalate esters, bisphenol analogues, and alternative plasticizer chemistries (Table 4, Figure 7). Applicability domain assessment was conducted for each compound using the leverage threshold recalculated for the top-20 feature space ($h^* = 0.423$) prior to interpreting the predicted values.

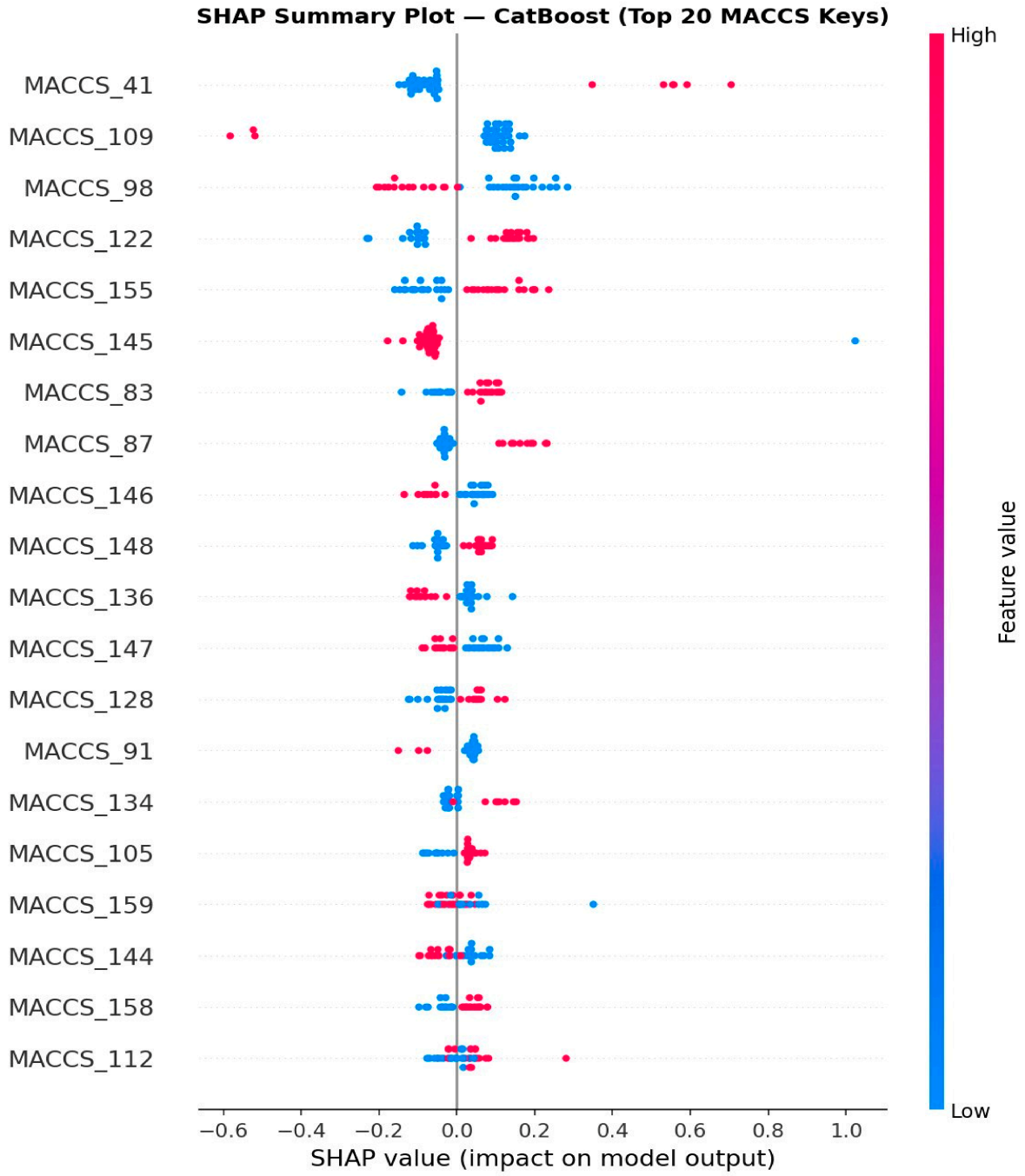


Figure 6. SHAP summary plot for the CatBoost model trained on the top 20 MACCS keys. Each point represents one test compound. The x-axis shows the SHAP value (impact on predicted pIC_{50}); colour indicates the feature value (red = present/high, blue = absent/low). MACCS_41 exerts the largest positive influence on predicted potency, confirming the importance of nitrogen-containing heterocyclic fragments in aromatase inhibition.

The dimensionality reduction achieved through permutation-based feature selection yielded a substantially more compact descriptor space, resulting in a tighter and geometrically well-defined applicability domain compared to the full 166-bit baseline. Consequently, eight of the ten plasticizers evaluated fell within the model's defined chemical space ($h \leq 0.423$), enabling reliable potency predictions for the majority of the evaluated compounds. Only bisphenol S (BPS; $h = 0.434$) and tributyl phosphate (TPBT; $h = 0.793$) marginally exceeded the leverage threshold, and their predictions should be treated as indicative rather than quantitative estimates.

Table 4. Predicted pIC_{50} values for representative plasticizers (CatBoost, top-20 MACCS features; h* = 0.423).

Abbrev.	Full Name	Predicted pIC_{50}	Predicted IC_{50} (nM)	Leverage (h)	Applicability Domain
DEHP	Bis(2-ethylhexyl) phthalate	8.22	6.0	0.169	Inside AD
DBP	Dibutyl phthalate	6.74	180.5	0.167	Inside AD
BBP	Benzyl butyl phthalate	4.13	74,329	0.175	Inside AD
DINP	Diisononyl phthalate	7.77	17.1	0.125	Inside AD
BPA	Bisphenol A	5.91	1226	0.052	Inside AD
BPS	Bisphenol S	5.70	2003	0.434	Outside AD
BPF	Bisphenol F	5.91	1226	0.052	Inside AD
DPHP	Dipentylhexyl phthalate	8.22	6.0	0.169	Inside AD
TPBT	Tributyl phosphate	6.85	141.7	0.793	Outside AD
ATBC	Acetyltributyl citrate	8.22	6.0	0.169	Inside AD

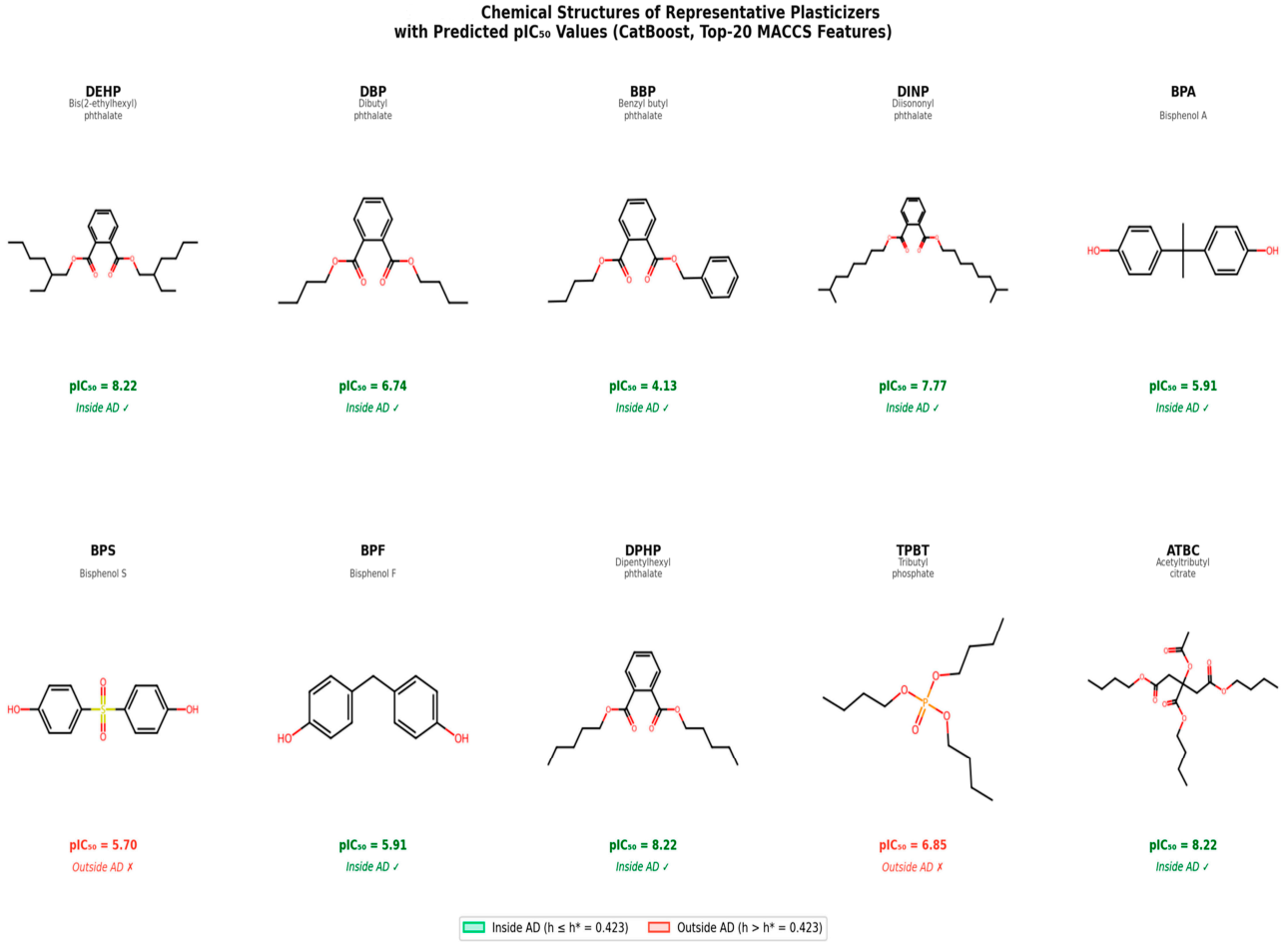


Figure 7. Chemical structures of ten representative plasticizers with predicted pIC_{50} values and applicability domain (AD) status. Eight compounds—DEHP (8.22), DINP (7.77), DPHP (8.22), ATBC (8.22), DBP (6.74), TPBT (6.85), BPA (5.91), BPF (5.91), and BBP (4.13)—fall within the AD ($h \leq 0.423$; green titles), while BPS (5.70) and TPBT (6.85) fall outside (red titles).

Among the compounds within the applicability domain, a clear stratification in predicted aromatase inhibitory potency emerged that is structurally coherent. The phthalate

esters DEHP, DINP, and DPHP received the highest predicted potency values (pIC_{50} = 8.22, 7.77, and 8.22 respectively; $IC_{50} \approx$ 6–17 nM), suggesting strong predicted inhibitory activity at nanomolar concentrations. DBP was predicted at pIC_{50} = 6.74 ($IC_{50} \approx$ 181 nM), representing moderate predicted potency, while BBP showed considerably weaker predicted activity (pIC_{50} = 4.13; $IC_{50} \approx$ 74,329 nM). The bisphenol compounds BPA and BPF received identical predicted pIC_{50} values of 5.91 ($IC_{50} \approx$ 1226 nM), consistent with their structural similarity and indicative of weak-to-moderate aromatase inhibitory activity. These values are broadly aligned with the range of experimental inhibitory activities reported for bisphenol compounds against CYP19A1 in the literature, where micromolar-range interactions with the aromatase active site have been documented [2,10]. The acetyltributyl citrate (ATBC) prediction of pIC_{50} = 8.22 should be interpreted with awareness that, despite falling within the leverage boundary, ATBC is structurally more distant from canonical aromatase inhibitor pharmacophores than the phthalate esters, and experimental validation would be required to substantiate this prediction.

From a structural perspective, the predicted potency hierarchy is mechanistically interpretable in the context of the feature importance findings. The phthalate esters, characterised by aromatic ester linkages and branched alkyl chains, share partial structural overlap with the nitrogen-containing and halogenated pharmacophoric elements identified as dominant in the SHAP and permutation importance analyses. By contrast, BBP's benzyl ester moiety introduces steric and electronic perturbations that appear to diminish predicted potency relative to the dialkyl phthalates, consistent with known structure–activity trends in this chemical class [7]. These findings collectively support the utility of the developed CatBoost framework as a first-tier screening tool for prioritising plasticizers with potential aromatase inhibitory activity, while reinforcing that experimental bioassay confirmation remains essential, particularly for compounds whose structural features diverge from the core training set pharmacophore.

4. Conclusions

This study developed and rigorously validated a machine learning QSAR framework for the prediction of aromatase (CYP19A1) inhibitory potency, expressed as pIC_{50}, using six regression algorithms paired with MACCS key molecular fingerprints. A systematically curated dataset of 187 compounds was assembled from ChEMBL (version 33), and a comprehensive validation strategy was implemented comprising an 80/20 stratified train/test split, 5-fold cross-validation, and Y-randomisation testing. A key methodological contribution of this work is the application of CatBoost permutation importance on the held-out test set for feature selection, identifying 20 MACCS keys from the initial 166-bit space whose selection substantially improved predictive generalisation across all algorithms. CatBoost trained on the top 20 features achieved the strongest test-set performance (R^2 = 0.693, RMSE = 0.794, MAE = 0.659) with the most stable cross-validation behaviour (CV R^2 = 0.062 $\pm$ 0.304), attributable to its ordered boosting mechanism and inherent regularisation properties. Y-randomisation returned an empirical p-value of < 0.01, confirming that model performance is grounded in genuine structure–activity relationships encoded in the selected descriptor space rather than statistical artefacts. SHAP and permutation importance analyses identified nitrogen-containing heterocyclic substructures (MACCS_41, MACCS_145) and halide-bearing fragments (MACCS_109) as the primary structural determinants of predicted inhibitory potency, findings that are pharmacologically coherent with the established haem-iron coordination and hydrophobic binding requirements of CYP19A1.

Application of the validated model to ten representative plasticizers demonstrated the practical utility of the framework. The transition to the top-20 feature space substan-

tially refined the applicability domain (h* = 0.423), enabling reliable predictions for eight of the ten plasticizers evaluated. The predicted potency values stratified meaningfully across structural classes: phthalate esters such as DEHP and DINP were predicted at nanomolar potency, BPA and BPF at the low micromolar range consistent with published experimental data, and BBP showed considerably weaker predicted activity. BPS and TPBT fell marginally outside the applicability domain and their predictions should be interpreted as exploratory. These results demonstrate that the model can function as a credible first-tier prioritisation tool for identifying plasticizers with potential endocrine disruption risk via aromatase inhibition, a capability of direct relevance to chemical risk assessment and the rational design of safer alternatives. Future work should prioritise the incorporation of experimentally characterised plasticizer-specific bioactivity data to extend the model's chemical coverage, exploration of extended connectivity fingerprints (ECFP4) and physicochemical descriptor sets, and systematic hyperparameter optimisation to further enhance predictive performance across the broader plasticizer chemical space.

Author Contributions: Conceptualization, I.L.M. and W.B.M.; methodology, I.L.M.; software, I.L.M.; validation, I.L.M., W.B.M. and N.R.; formal analysis, I.L.M.; investigation, I.L.M.; resources, I.L.M.; data curation, I.L.M.; writing—original draft preparation, I.L.M.; writing—review and editing, I.L.M., W.B.M., N.R. and J.-N.O.; visualization, I.L.M.; supervision, W.B.M., N.R. and J.-N.O.; project administration, W.B.M., N.R. and J.-N.O.; funding acquisition, W.B.M., N.R. and J.-N.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation (PMDS230627122873) and Sefako Makgatho Health Sciences University.

Institutional Review Board Statement: This study utilized publicly available data obtained from the ChEMBL database (Subject ID: ChEMBL1978). No human participants or experimental animals were directly involved in this research. Therefore, ethical approval and informed consent were not required.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analysed in this study were obtained from the publicly available ChEMBL database.

Acknowledgments: This work is supported by the National Research Foundation and the University of Sefako Makgatho Health Sciences.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, W.; Fang, F.; Zhu, W.; Chen, Z.-J.; Du, Y.; Zhang, J. Bisphenol A and ovarian reserve among infertile women with polycystic ovarian syndrome. *Int. J. Environ. Res. Public Health* **2017**, *14*, 18.
2. Huo, W.; Xia, W.; Wan, Y.; Zhang, B.; Zhou, A.; Zhang, Y.; Huang, K.; Zhu, Y.; Wu, C.; Peng, Y. Maternal urinary bisphenol A levels and infant low birth weight: A nested case–control study of the Health Baby Cohort in China. *Environ. Int.* **2015**, *85*, 96–103. [CrossRef]
3. Komarowska, M.D.; Grubczak, K.; Czerniecki, J.; Hermanowicz, A.; Hermanowicz, J.M.; Debek, W.; Matuszczak, E. Identification of the Bisphenol A (BPA) and the Two Analogues BPS and BPF in Cryptorchidism. *Front. Endocrinol.* **2021**, *12*, 694669. [CrossRef]
4. Mahlangu, W.B.; Maseko, B.R.; Mongadi, I.L.; Makhubela, N.; Ncube, S. Quantitative analysis and health risk assessment of bisphenols in selected canned foods using the modified QuEChERS method coupled with gas chromatography-mass spectrometry. *Food Packag. Shelf Life* **2023**, *37*, 101078. [CrossRef]
5. Lehmler, H.-J.; Liu, B.; Gadogbe, M.; Bao, W. Exposure to bisphenol A, bisphenol F, and bisphenol S in US adults and children: The national health and nutrition examination survey 2013–2014. *ACS Omega* **2018**, *3*, 6523–6532. [CrossRef]
6. Qadeer, A.; Kirsten, K.L.; Ajmal, Z.; Jiang, X.; Zhao, X. Alternative plasticizers as emerging global environmental and health threat: Another regrettable substitution? *Environ. Sci. Technol.* **2022**, *56*, 1482–1488. [CrossRef]

7. Rochester, J.R.; Bolden, A.L. Bisphenol S and F: A systematic review and comparison of the hormonal activity of bisphenol A substitutes. *Environ. Health Perspect.* **2015**, *123*, 643–650. [CrossRef] [PubMed]
8. Struzina, L.; Castro, M.A.P.; Kubwabo, C.; Siddique, S.; Zhang, G.; Fan, X.; Tian, L.; Bayen, S.; Aneck-Hahn, N.; Bornman, R. Occurrence of legacy and replacement plasticizers, bisphenols, and flame retardants in potable water in Montreal and South Africa. *Sci. Total Environ.* **2022**, *840*, 156581. [CrossRef]
9. Di Nardo, G.; Zhang, C.; Marcelli, A.G.; Gilardi, G. Molecular and structural evolution of cytochrome P450 aromatase. *Int. J. Mol. Sci.* **2021**, *22*, 631. [CrossRef] [PubMed]
10. Yoshimoto, F.K.; Guengerich, F.P. Mechanism of the third oxidative step in the conversion of androgens to estrogens by cytochrome P450 19A1 steroid aromatase. *J. Am. Chem. Soc.* **2014**, *136*, 15016–15025. [CrossRef] [PubMed]
11. Turner, K.; Macpherson, S.; Millar, M.; Mcneilly, A.; Williams, K.; Cranfield, M.; Groome, N.; Sharpe, R.; Fraser, H.; Saunders, P. Development and validation of a new monoclonal antibody to mammalian aromatase. *J. Endocrinol.* **2002**, *172*, 21–30. [CrossRef] [PubMed]
12. Hackett, J.C.; Brueggemeier, R.W.; Hadad, C.M. The final catalytic step of cytochrome P450 aromatase: A density functional theory study. *J. Am. Chem. Soc.* **2005**, *127*, 5224–5237. [CrossRef] [PubMed]
13. Caldwell, G.W.; Yan, Z.; Lang, W.; Masucci, J.A. The IC50 concept revisited. *Curr. Top. Med. Chem.* **2012**, *12*, 1282–1290. [CrossRef]
14. Geisler, J. Differences between the non-steroidal aromatase inhibitors anastrozole and letrozole–of clinical importance? *Br. J. Cancer* **2011**, *104*, 1059–1066. [CrossRef] [PubMed]
15. Kijima, I.; Itoh, T.; Chen, S. Growth inhibition of oestrogen receptor-positive and aromatase-positive human breast cancer cells in monolayer and spheroid cultures by letrozole, anastrozole, and tamoxifen. *J. Steroid Biochem. Mol. Biol.* **2005**, *97*, 360–368. [CrossRef]
16. Soares, T.A.; Nunes-Alves, A.; Mazzolari, A.; Ruggiu, F.; Wei, G.-W.; Merz, K. The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) studies propelled by the surge of machine learning methods. *J. Chem. Inf. Model.* **2022**, *62*, 5317–5320.
17. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Towards understanding aromatase inhibitory activity via QSAR modeling. *EXCLI J.* **2018**, *17*, 688–708.
18. Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E.J.; Blackshaw, J.; Corbett, S.; De Veij, M.; Ioannidis, H.; Lopez, D.M.; Mosquera, J.F.; et al. The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **2024**, *52*, D1180–D1192. [CrossRef]
19. Landrum, G.A. *RDKit: Open-Source Cheminformatics Software*, version 2023.09.1; Zenodo: Geneva, Switzerland, 2023. [CrossRef]
20. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701. [CrossRef]
21. Dearden, J.C.; Cronin, M.T.D.; Kaiser, K.L.E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20*, 241–266. [CrossRef]
22. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6638–6648.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.