Journal of *Xenobiotics*

MDPI

Article

QSAR Models for Predicting Oral Bioavailability and Volume of Distribution and Their Application in Mapping the TK Space of Endocrine Disruptors

Guillaume Ollitrault [1], Marco Marzo [2], Alessandra Roncaglioni [2], Emilio Benfenati [2], Olivier Taboureau [1,*] and Enrico Mombelli [3]

[1] Inserm U1133, CNRS UMR 8251, Université de Paris Cité, 75013 Paris, France; guillaume.ollitrault@inserm.fr
[2] Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, 20156 Milano, Italy; marco.marzo@marionegri.it (M.M.); alessandra.roncaglioni@marionegri.it (A.R.); emilio.benfenati@marionegri.it (E.B.)
[3] Institut National de l'Environnement Industriel et des Risques (INERIS), 60550 Verneuil en Halatte, France; enrico.mombelli@ineris.fr
* Correspondence: olivier.taboureau@u-paris.fr

Abstract

Toxicokinetic (TK) properties are essential in the framework of chemical risk assessment and drug discovery. Specifically, a TK profile provides information about the fate of chemicals in the human body. In this context, Quantitative Structure–Activity Relationship (QSAR) models are convenient computational tools for predicting TK properties. Here, we developed QSAR models to predict two TK properties: oral bioavailability and volume of distribution at steady state (VD_{ss}). We collected and curated two large sets of 1712 and 1591 chemicals for oral bioavailability and VD_{ss}, respectively, and compared regression and classification (binary and multiclass) models with the application of several machine learning algorithms. The best predictive performance of the models for regression (R) prediction was characterized by a Q^2_{F3} of 0.34 with the R-CatBoost model for oral bioavailability and a geometric mean fold error (GMFE) of 2.35 with the R-RF model for VD_{ss}. The models were then applied to a list of potential endocrine-disrupting chemicals (EDCs), highlighting chemicals with a high probability of posing a risk to human health due to their TK profiles. Based on the results obtained, insights into the structural determinants of TK properties for EDCs are further discussed.

Keywords: QSAR; oral bioavailability; volume of distribution; endocrine-disrupting chemicals; Toxicokinetics

Academic Editor: Elisa Cairrao

Received: 4 August 2025
Revised: 11 September 2025
Accepted: 17 September 2025
Published: 15 October 2025

Citation: Ollitrault, G.; Marzo, M.; Roncaglioni, A.; Benfenati, E.; Taboureau, O.; Mombelli, E. QSAR Models for Predicting Oral Bioavailability and Volume of Distribution and Their Application in Mapping the TK Space of Endocrine Disruptors. *J. Xenobiot.* **2025**, *15*, 166. https://doi.org/10.3390/jox15050166

1. Introduction

In the realm of drugs, pharmacokinetics (PK) refers to the characterization of the absorption, distribution, metabolism, and excretion of xenobiotics in an organism [1]. Toxicokinetics (TK) is closely related to pharmacokinetics (PK), as it involves the generation of PK data, either as part of nonclinical toxicity studies or through dedicated supportive studies to evaluate systemic exposure. Such analyses are largely used in pharmaceutical and chemical industries, as they are critical for gaining insights into the TK propensities of potential drug candidates and for assessing risks associated with environmental chemicals [2,3].

To limit the costs of such experiments and still provide relevant information for decision-makers in the field of drug discovery and chemical risk assessment, in silico models capable of predicting key TK/PK properties, notably oral bioavailability and volume of distribution (VD), are commonly developed for initial estimation [4].

Oral bioavailability characterizes the fraction of an orally administered drug that reaches the systemic circulation (F%). It is calculated based on the relationship between plasma chemical concentration and time after administration. Oral bioavailability is defined as the percentage of the dose area under the curve of the chemical concentration in the plasma after oral administration, divided by the dose area under the curve of the concentration of the drug in the plasma after intravenous administration [5]. This comparison yields information about the proportion of chemicals reaching the bloodstream since intravenous administration circumvents the digestive system and first-pass metabolism. High oral bioavailability can result in exposure to toxic compounds after intake, and low oral bioavailability for drugs can increase the required dose, with the associated risk of toxicity through accumulation and metabolites [6].

The volume of distribution (VD) measures the ability of a chemical to remain in plasma or to redistribute to other tissue compartments. VD is computed by considering the amount of a chemical in the body divided by the plasma concentration of the same chemical [7]. In the field of drug discovery, having a priori knowledge about VD assists in optimizing drug therapies, avoiding undesirable effects, and proposing effective treatments. Specifically, at a constant clearance rate, a chemical with a high VD will have a longer elimination half-life than one with a low VD [8], since the former will persist in tissues while being slowly released into the bloodstream. Therefore, knowing the VD of environmental chemicals is also important in the field of chemical risk assessment since these chemicals might remain longer in tissues, which could lead to accumulation in the human body and result in toxicity, especially for lipophilic drugs [6]. Different VD-related terms are commonly used, with the volume of distribution at steady state (VD_{ss}) generally being the most relevant, as it is used to determine the VD associated with the steady-state dosing of the chemical. It is calculated during the phase called "steady state", when the distribution and elimination phases are equal [7].

Several computational studies have been conducted to predict oral bioavailability [9–16] and VD_{ss}, and many have used QSAR models [17–21]. Most existing models have focused on oral bioavailability using classification approaches, while regression models have been primarily developed for VD_{ss}, notably using Lombardo et al.'s dataset [17]. In our work, we combined datasets from multiple sources, including a newly developed dataset from Liu et al. [22].

In this context, we decided to collect a large dataset of chemicals and to develop different modeling algorithms for regression, binary-class, and multiclass prediction for oral bioavailability and VD_{ss}.

The most relevant models were then used to assess the TK properties of potential endocrine-disrupting chemicals (EDCs). The focus on this category of chemicals is motivated by the fact that EDCs can disrupt the endocrine system and cause cancer, metabolic disorders, neurocognitive functions, infertility, immune diseases, and allergies [23–27] by interfering with the estrogen, androgen, and thyroid hormone receptors, exerting steroidogenesis (ER, AR, and TR)-mediated effects [28]. Thus, predicting potential EDCs with high oral bioavailability and high VD_{ss} could be relevant for regulatory purposes.

To complement this work, we also applied an existing QSAR model to predict the elimination half-lives ($t_{1/2}$) of EDCs. The elimination half-life is a key toxicokinetic parameter that reflects the time required for the concentration of a chemical in the body to decrease by half. This feature is crucial for assessing a compound's persistence, bioaccumulation

potential, and dosing frequency, making it an important factor in risk assessment and regulatory decision-making [29,30].

Finally, this large-scale analysis provides insights into the structural features that might be important in the determination of TK for EDCs, which is further discussed below.

2. Results

2.1. Data Distribution

Starting from the chemicals with experimentally known F% and VD_{ss}, multiple datasets were designed. The number of chemicals in each dataset is reported in Table 1.

Table 1. Number of chemicals used to develop QSAR models, according to the modeling algorithms, for oral bioavailability and VD_{ss}.

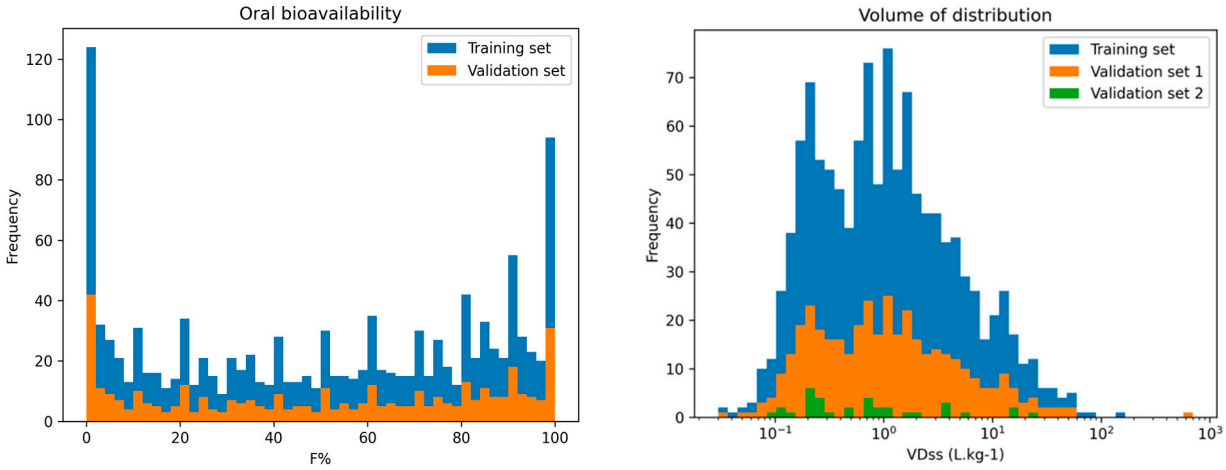
Endpoint	Dataset	Modeling Algorithm	Number of Chemicals
Oral bioavailability	Training	Regression	1213
		Classification (50% threshold)	1307
		Binary classification (30% and 60% thresholds)	1244
	Validation	Regression/binary classification/multiclass classification	405
VD_{ss}	Training	Regression/binary classification/multiclass classification	1167
	Validation 1		390
	Validation 2		34

The work described herein relied on three datasets for training models to predict oral bioavailability. The first dataset contained 1213 chemicals and was used to train regression models. The second dataset was composed of 1307 chemicals and was used to train classification models with a 50% dichotomizing threshold. The third dataset consisted of 1244 chemicals and was used to train multiclass models. All models trained on the three datasets were then evaluated on a common set of 405 chemicals with known F% values.

For the VD_{ss} analysis, a single dataset containing 1167 chemicals was used to train regression and multiclass classification models. Two validation sets, one with 390 chemicals and the other with 34, were used. The first set was used to assess the overall performance of the trained model both with and without applying applicability domains, whereas the second set was used to compare the model's predictive performance with that of published QSAR models for the same endpoints, given that it is a set of chemicals commonly used to compare the precision of QSAR models in the literature.

2.1.1. Oral Bioavailability Data

The distributions of F% for the training and validation sets cover the complete endpoint range while having a similar shape, and therefore, they are suitable for model evaluation and training (Figure 1a). Indeed, the bioavailability values span the entire range from 0% to 100%. The distribution exhibits peaks at 0% and 100% bioavailability. This characteristic could be due to the limitations of oral bioavailability testing methods, as discussed by Aungst et al. [31]. The presence of many chemicals associated with 0% or 100%, with few in-between values, likely introduces a bias where the model yields correct predictions for the majority classes while displaying poor performance for intermediate values.



(**a**) (**b**)

Figure 1. (**a**) Histograms of the distribution of oral bioavailability for the training set and the validation set for all chemicals with continuous F% values. (**b**) Distribution of the values characterizing VD_{ss} for the training set and validation sets 1 and 2.

2.1.2. Volume of Distribution Data

Figure 1b shows the distribution of VD_{ss} values across the training set, validation set 1, and validation set 2. To address the skewed nature of the VD_{ss} distribution (from 0.035 $L \cdot kg^{-1}$ to 700 $L \cdot kg^{-1}$) and facilitate model convergence, the dependent variable was logarithmically transformed, in base e, when regression models were applied. The distribution of VD_{ss} values across the training set, validation set 1, and validation set 2 is depicted in Figure 1b. We can observe that all three datasets exhibit comparable distributions across this range, ensuring coverage of VD_{ss} values for model training and evaluation.

2.1.3. Chemical Space

The chemical space covered by the chemical sets was characterized using a Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) representation [32], facilitating a comprehensive exploration of the molecular landscape and enabling insightful analysis of the distribution of chemicals (Supplementary Figure S1).

UMAP on the oral bioavailability dataset (Supplementary Figure S1a) shows the distribution of chemicals while accounting for their F% values. The plot reveals that most points are distributed all around the two axes and exhibit a wide range of F% values, highlighting the difficulty of finding patterns between F% values and chemical similarity.

The UMAP representation of the $\ln(VD_{ss})$ values (Supplementary Figure S1b) projects the high-dimensional VD_{ss} dataset onto a two-dimensional map. The plot highlights the range of VD_{ss} values, with higher values on the left and lower values on the right, illustrating the relationship between chemical similarity and VD_{ss} values.

These observed patterns support the pertinence of using machine learning models for F% and VD_{ss} prediction. Machine learning algorithms can potentially learn effective predictive models that capture the diverse landscapes observed in these datasets.

2.2. Predictive Performance

2.2.1. Oral Bioavailability Performance

Multiple models were trained and evaluated to predict oral bioavailability. Regression models were evaluated for the prediction of continuous values, while for binary-class and multiclass prediction, we imposed 50% and 30–60% thresholds. All models were evaluated using dedicated metrics.

From the 1826 molecular descriptors computed with Mordred, the most relevant ones were selected using the VSURF algorithm. This resulted in the selection of 66 molecular descriptors for a Topliss ratio (number of training chemicals per molecular descriptor) of 18:1 for the regression model, 59 molecular descriptors (Topliss ratio of 26:1) for binary classification prediction with a 50% threshold, and 70 molecular descriptors (Topliss ratio of 23:1) for multiclass prediction with 30–60% thresholds, with the Topliss ratio largely in compliance with the recommended threshold (>5) to avoid overfitting. Then, these selected molecular descriptors were utilized as input features to train the CatBoost, XGBoost, and RF models for predictive modeling.

The predictive performance of the algorithms was evaluated across regression (R), binary classification (BC), and multiclass classification (MC) tasks, with the regression task further assessed for its ability to facilitate classification-based predictions for the training and validation sets.

As the majority of the models developed showed high performance values on the training sets (Supplementary Tables S2a, S3a and S4a), we evaluated the performance using five-fold cross-validation on training sets in order to select the best models. More precisely, models characterized by the highest mean Q^2_{F3}, BA, and macro-BA for, respectively, regression, binary classification, and multiclass classification (for internal validation), were selected. Ultimately, the R-CatBoost, BC-CatBoost, and MC-CatBoost models were retained as the best models since they were characterized by the highest Q^2_{F3}, BA, and macro-BA (0.34 ± 0.05, 0.74 ± 0.02, 0.69 ± 0.02, respectively) (Table 2).

Table 2. Performance obtained for the QSAR models in predicting oral bioavailability on the validation set. The predictive performance of the algorithms was evaluated across regression (R), binary classification (BC), and multiclass classification (MC) tasks. NA means not applicable.

Metric	Performance for Regression (R)	Performance for Binary Classification (BC)	Performance for Multiclass Classification (MC)	Cross-Validation (CV) Performance for Regression (R)	CV Performance for Binary Classification (BC)	CV Performance for Multiclass Classification (MC)
		Validation Set			CV	
Model	**R-CatBoost**	**BC-CatBoost**	**MC-CatBoost**	**R-CatBoost**	**BC-CatBoost**	**MC-CatBoost**
Regression metrics						
RMSE	25.86	NA		27.71 ± 0.98		
R^2	0.42	NA	NA	0.38 ± 0.04	NA	NA
MAE	20.09	NA	NA	20.90 ± 0.82	NA	NA
MedAE	15.92	NA	NA	17.01 ± 1.11	NA	NA
Q^2_{F3}	0.39	NA	NA	0.34 ± 0.05	NA	NA
Binary classification metrics						
Sensitivity	0.78	0.79	NA	0.75 ± 0.03	0.78 ± 0.03	NA
Specificity	0.76	0.68	NA	0.72 ± 0.03	0.69 ± 0.04	NA
Balanced accuracy	0.77	0.74	NA	0.74 ± 0.02	0.74 ± 0.02	NA
Multiclass classification metrics						
Sensitivity (<30%)	0.46	NA	0.67	0.45 ± 0.05	NA	0.64 ± 0.05
Specificity (<30%)	0.91	NA	0.86	0.93 ± 0.03	NA	0.83 ± 0.03
Balanced accuracy (<30%)	0.68	NA	0.77	0.63 ± 0.02	NA	0.74 ± 0.02
Sensitivity [30–60%]	0.58	NA	0.25	0.69 ± 0.02	NA	0.31 ± 0.05
Specificity [30–60%]	0.63	NA	0.89	0.63 ± 0.03	NA	0.88 ± 0.02
Balanced accuracy [30–60%]	0.60	NA	0.57	0.63 ± 0.03	NA	0.60 ± 0.03
Sensitivity (>60%)	0.63	NA	0.83	0.63 ± 0.04	NA	0.79 ± 0.03
Specificity (>60%)	0.84	NA	0.67	0.84 ± 0.03	NA	0.70 ± 0.03
Balanced accuracy (>60%)	0.74	NA	0.75	0.74 ± 0.02	NA	0.74 ± 0.02
Macro sensitivity	0.56	NA	0.58	0.57 ± 0.03	NA	0.58 ± 0.02
Macro specificity	0.79	NA	0.81	0.80 ± 0.01	NA	0.81 ± 0.01

Table 2. *Cont.*

Metric	Performance for Regression (R)	Performance for Binary Classification (BC)	Performance for Multiclass Classification (MC)	Cross-Validation (CV) Performance for Regression (R)	CV Performance for Binary Classification (BC)	CV Performance for Multiclass Classification (MC)
		Validation Set			CV	
Model	**R-CatBoost**	**BC-CatBoost**	**MC-CatBoost**	**R-CatBoost**	**BC-CatBoost**	**MC-CatBoost**
Macro balanced accuracy	0.68	NA	0.70	0.68 ± 0.02	NA	0.69 ± 0.02
Micro sensitivity	0.56	NA	0.64	0.57 ± 0.03	NA	0.63 ± 0.02
Micro specificity	0.78	NA	0.82	0.79 ± 0.01	NA	0.82 ± 0.01

For the validation set, the R-CatBoost regression algorithm achieved an R^2 of 0.43 and a Q^2_{F3} of 0.39 (Table 2, Supplementary Table S2a). Furthermore, the mean absolute error (MAE) is reported at 20.09 (F%) within the range of 0 to 100 (F%). The RMSE is also significant, with a value of 25.86 (F%). Absolute F% errors of 10 and 20 are illustrated in Supplementary Figure S3. According to Wang et al., the RMSE of experimental measurements of oral bioavailability is 14.5 (F%) [33], which might explain this high RMSE.

We categorized the outcome predictions from the developed R-CatBoost model into two classes: high (greater than 50%) and low (less than 50%) oral bioavailability. We then evaluated the performance for binary classification using the 50% threshold, resulting in a BA of 0.77 (Table 2, Supplementary Table S3a). In comparison, the best model trained on binary data, where values are dichotomized into 1 (greater than 50%) and 0 (less than 50%), showed a lower BA, with the BC-CatBoost classification method achieving a BA of 0.74 (Supplementary Table S3a).

We applied the same processing for multiclass classification; we categorized outcome predictions from R-CatBoost regression into three classes: low (less than 30%), medium (higher than 30% and less than 60%), and high (higher than 60%) oral bioavailability. We then evaluated the performance for multiclass classification using the 30% and 60% thresholds, resulting in a lower macro-BA of 0.67 compared to the multiclass model, MC-CatBoost, which achieved a macro-BA of 0.70. The analysis of predictive performance under the 30–60% thresholds (Table 2, Supplementary Table S4a) further highlights notable trends and disparities among various machine learning approaches in multiclass prediction. On the validation set, the MC-CatBoost model trained for multiclass prediction achieved a BA of 0.77 for the <30% class and of 0.75 for the >60% class. However, these models had low reliability when predicting the intermediate class (between 30% and 60%), showing a lower BA of 0.57, alongside a pronounced inability to accurately identify chemicals in this range, exemplified by an SE of 0.25.

The R-CatBoost regression model, while exhibiting lower performance compared to the double-threshold MC-CatBoost classification model, offers superior versatility and effectiveness in predicting medium-F% chemicals. It achieved a BA of 0.60 and an SE of 0.58 for the medium-F% class, demonstrating its utility in addressing the complexities of multiclass prediction tasks. These results emphasize the importance of methodology selection, with regression models proving particularly advantageous for medium-class prediction.

2.2.2. Volume of Distribution Performance

Multiple ML models were trained and evaluated for their robustness in predicting VD_{ss}. Regression models were evaluated for the prediction of continuous values and for dichotomous and multiclass predictions with 1 $L \cdot kg^{-1}$ and 0.6 $L \cdot kg^{-1}$–5 $L \cdot kg^{-1}$ categorization thresholds. The models were evaluated using dedicated metrics.

Molecular descriptor selection utilizing the VSURF algorithm on Mordred molecular descriptors yielded a subset of 26 molecular descriptors for a Topliss ratio (number of

training chemicals per molecular descriptor) of 45:1, in compliance with the recommended threshold (>5) to avoid overfitting. These selected molecular descriptors were used as input to train the CatBoost, XGBoost, and RF models.

In order to select the best models, we only considered the performance obtained in five-fold cross-validation on training sets. More precisely, the best models corresponded to those characterized by the lowest mean GMFE, the highest BA, and the highest macro-BA for, respectively, the regression, binary classification, and multiclass models (for internal validation). According to this logic, the R-RF, BC-Chemprop, and MC-Chemprop models were retained as the best models since they were characterized by the lowest GMFE, highest BA, and macro-BA (2.19 ± 0.08, 0.78 ± 0.02, and 0.73 ± 0.02, respectively) (Supplementary Tables S5b, S6b and S7b). These models performed well on training data, with a GMFE below 2 (Supplementary Tables S5a, S6a and S7a).

For validation set 1, the R-RF regression algorithm achieved a GMFE of 2.35 (Supplementary Table S4a), indicating that the model can be regarded as sufficiently precise [34]. The R-RF regression model was able to predict VD_{ss} values mostly within 2-fold to 3-fold errors (Supplementary Figure S4).

From the developed R-RF model, we categorized the outcome predictions into two classes: high (greater than 1 $L \cdot kg^{-1}$) and low (less than 1 $L \cdot kg^{-1}$) VD_{ss}. We then evaluated the performance for binary classification using the 1 $L \cdot kg^{-1}$ threshold, resulting in a BA of 0.75, showing comparable performance to that of the best model, BC-Chemprop trained on binary data, where values are dichotomized into 1 (greater than 1 $L \cdot kg^{-1}$) and 0 (less than 1 $L \cdot kg^{-1}$), which achieved a BA of 0.76 (Supplementary Table S6a).

We applied the same processing for multiclass classification; we categorized the outcome predictions from R-RF regression into three classes: low (less than 0.6 $L \cdot kg^{-1}$), medium (higher than 0.6 $L \cdot kg^{-1}$ and less than 5 $L \cdot kg^{-1}$), and high (higher than 5 $L \cdot kg^{-1}$) VD_{ss}. We then evaluated the performance for multiclass classification using the 0.6 $L \cdot kg^{-1}$ and 5 $L \cdot kg^{-1}$ thresholds, resulting in a lower macro-BA of 0.68 (Table 3) compared to the best-performing multiclass model, MC-Chemprop, which achieved a macro-BA of 0.72. The analysis of predictive performance under the 0.6–5 $L \cdot kg^{-1}$ threshold (Table 3) did not reveal notable trends, as all classes had a BA greater than 0.60 for all algorithms (Supplementary Table S7a).

Table 3. Predictive performance of the QSAR models for the prediction of VD_{ss} as a function of validation set 1. The predictive performance of the algorithms was evaluated across regression (R), binary classification (BC), and multiclass classification (MC) tasks.

Metric	Regression Model Performance	Classification Model Performance	Multiclass Classification Model Performance	CV Regression Model Performance	CV Classification Model Performance	CV Multiclass Classification Model Performance
		Validation Set 1			CV	
Model	**R-RF**	**BC-Chemprop**	**MC-Chemprop**	**R-RF**	**BC-Chemprop**	**MC-Chemprop**
Regression metrics						
GMFE	2.35	NA	NA	2.19 ± 0.08	NA	NA
Binary Classification metrics						
Sensitivity	0.79	0.77	NA	0.79 ± 0.03	0.73 ± 0.06	NA
Specificity	0.71	0.75	NA	0.75 ± 0.03	0.83 ± 0.04	NA
Balanced accuracy	0.75	0.76	NA	0.77 ± 0.02	0.78 ± 0.03	NA
Multiclass classification metrics						
Sensitivity (<0.6)	0.62	NA	0.68	0.66 ± 0.04	NA	0.71 ± 0.05
Specificity (<0.6)	0.91	NA	0.89	0.90 ± 0.02	NA	0.87 ± 0.03
Balanced accuracy (<0.6)	0.76	NA	0.78	0.78 ± 0.02	NA	0.79 ± 0.03
Sensitivity [0.6–5]	0.82	NA	0.76	0.83 ± 0.03	NA	0.76 ± 0.05

Table 3. *Cont.*

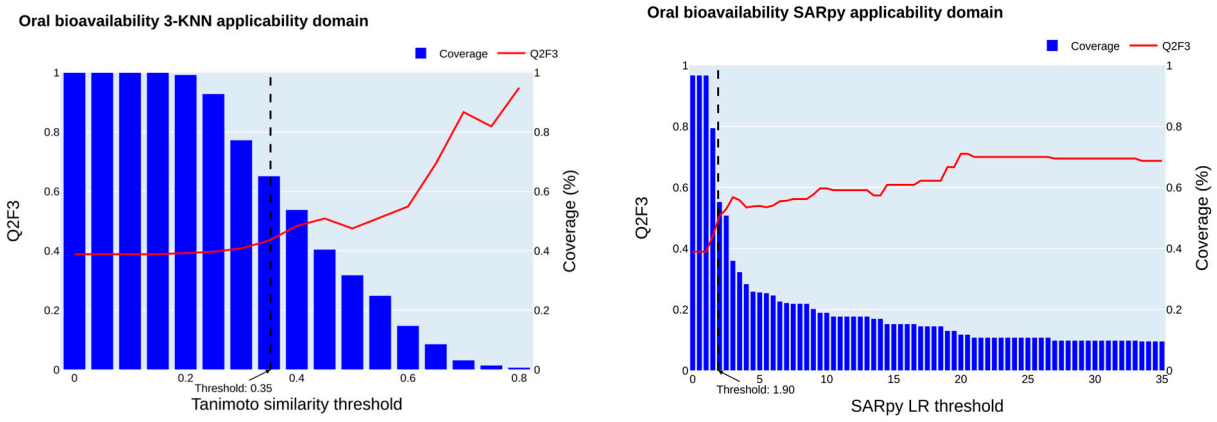
Metric	Regression Model Performance	Classification Model Performance	Multiclass Classification Model Performance	CV Regression Model Performance	CV Classification Model Performance	CV Multiclass Classification Model Performance
		Validation Set 1			CV	
Model	**R-RF**	**BC-Chemprop**	**MC-Chemprop**	**R-RF**	**BC-Chemprop**	**MC-Chemprop**
Specificity [0.6–5]	0.51	NA	0.63	0.57 ± 0.04	NA	0.66 ± 0.05
Balanced accuracy [0.6–5]	0.67	NA	0.70	0.70 ± 0.02	NA	0.71 ± 0.03
Sensitivity (>5)	0.22	NA	0.45	0.32 ± 0.06	NA	0.42 ± 0.09
Specificity (>5)	0.97	NA	0.94	0.97 ± 0.01	NA	0.94 ± 0.02
Balanced accuracy (>5)	0.60	NA	0.69	0.64 ± 0.03	NA	0.68 ± 0.04
Macro sensitivity	0.56	NA	0.63	0.60 ± 0.03	NA	0.63 ± 0.03
Macro specificity	0.80	NA	0.82	0.81 ± 0.01	NA	0.82 ± 0.02
Macro balanced accuracy	0.68	NA	0.72	0.71 ± 0.02	NA	0.73 ± 0.02
Micro sensitivity	0.65	NA	0.68	0.68 ± 0.02	NA	0.69 ± 0.03
Micro specificity	0.83	NA	0.84	0.84 ± 0.01	NA	0.84 ± 0.01
Micro balanced accuracy	0.74	NA	0.76	0.76 ± 0.02	NA	0.77 ± 0.02

The R-RF model AD was further explored, and the model was used to map EDCs.

2.3. Applicability Domain

2.3.1. Oral Bioavailability Applicability Domain

The applicability domain of the regression model (R-CatBoost was assessed according to the best mean Q^2_{F3} of the 50 iterations of 5f-CV) was assessed using a three-nearest neighbor approach on the validation set. The same plot reports the Q^2_{F3} performance and the coverage according to different Tanimoto thresholds for the R-CatBoost regression models (Figure 2a). This model showed good overall performance in predicting oral bioavailability for low, medium, and high categories. As the applicability domain narrows, the validation set comprises more structurally similar compounds to the training set, and our models exhibit enhanced performance.



(**a**) (**b**)

Figure 2. The effect of different definitions of applicability domains on coverage and predictive performance. (**a**) The 3-NN Tanimoto AD. The evolution of Q^2_{F3} performance between observed and predicted values (red) on the validation set, between predicted and real values according to different Tanimoto thresholds ranging from 0 to 1. The evolution of the validation set coverage is plotted as blue bars. The model tested is the CatBoost regression method predicting F% values. (**b**) The SARpy LR AD. The evolution of Q^2_{F3} performance (red) on the validation set according to different Log ratio thresholds relative to the structural alert associated with query chemicals. The evolution of the validation set coverage is plotted as blue bars.

This improvement stems from the models' ability to effectively recognize and learn the inherent patterns within the data. However, at higher threshold levels, occasional declines in R^2 performance are observed. These fluctuations arise due to certain compounds being inaccurately predicted despite their structural resemblance to those in the training set. Additionally, as the number of compounds used for performance evaluation decreases, the uncertainty in performance metrics increases. Assessing performance based on a small dataset introduces variability, which can compromise the reliability and robustness of the models. While restricting the applicability domain can enhance performance, it is crucial to maintain a balance between predictive accuracy and the number of retained compounds to ensure the validity of the models. Here, we considered a minimum coverage of 60% of chemicals in the validation set retained, corresponding to a Q^2_{F3} of 0.46.

Finally, we considered a Tanimoto threshold of 0.35 when applying the threshold formula Dc = <y> + Z × sigma, with <y> equal to 0.35, Z equal to 0.5, and a sigma of 0.14. This threshold resulted in a Q^2_{F3} improvement from 0.39 to 0.43 and an MAE decrease from 20.09 to 18.9 with a coverage of 65%.

We explored the use of the Log ratio (LR) given by the MC-SARpy multiclass model to define the applicability domain. We plotted Q^2_{F3} by varying the LR threshold from 0 to the maximum values of LR (infinite values transformed to maximum LR) alongside the size of the retained validation set (Figure 2b). Q^2_{F3} increases as the thresholds are raised. Structural fragments defined by the MC-SARpy model (Supplementary Table S8) can be employed to provide insights into the reliability of predictions and identify significant structural features that move chemicals toward either high or low F% values.

When we use SARpy to define the applicability domain and consider a threshold corresponding to a coverage of 65% (LR of 1.90), as we did when analyzing the applicability domain defined by the k-nearest neighbor approach, we obtain a Q^2_{F3} of 0.46. This predictive performance is slightly better than that obtained with the k-nearest neighbor method and can be used as the applicability domain definition to improve the model's performance. Both approaches can be used together to define the AD, each providing deeper insight into the prediction.

2.3.2. Volume of Distribution Applicability Domain

The applicability domain of the best regression model (R-RF was assessed according to the best mean GMFE of the 50 iterations of 5f-CV) was explored using a three-nearest neighbor approach on the validation set. In the same plot, the GMFE performance and the coverage of chemicals retained according to varying Tanimoto thresholds for the R-RF regression model are illustrated in Figure 3a.

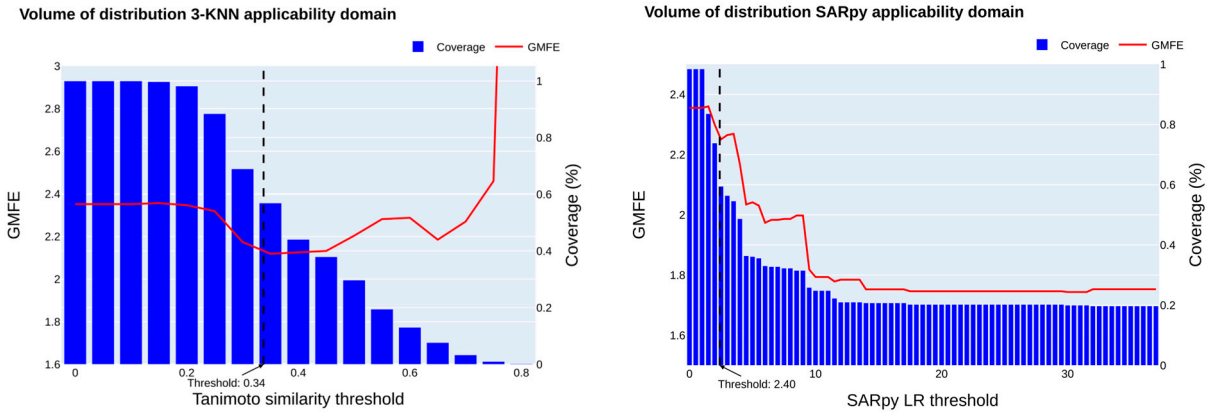
As the threshold is further increased, the GMFE stops decreasing and starts increasing, as was seen in the QSAR model for oral bioavailability. We considered a minimum coverage of 60% of chemicals in the validation set retained.

Finally, we considered a Tanimoto threshold of 0.34 when applying the threshold formula Dc = <y> − Z × sigma, with <y> equal to 0.42, Z equal to 0.5, and a sigma of 156. This threshold improved the GMFE from 2.35 to 2.17, becoming closer to 2, with a coverage of 61%.

We used the Log ratio (LR) given by the MC-SARpy multiclass model as an applicability domain definition. We plotted the GMFE by varying the LR threshold from 0 to the maximum values of LR (infinite values transformed to maximum LR), alongside the effective data retained in the validation set (Figure 3b).

The GMFE decreases as the thresholds are increased. When a threshold corresponding to a coverage of 63% is considered (LR of 2.40), similarly to what is described for the k-nearest neighbor approach, a GMFE equal to 2.24 is observed. SARpy structural fragments

defined by the MC-SARpy model can be used (Supplementary Table S9) to provide insights into the reliability of predictions and identify significant structural features that modulate chemical activity toward either high or low VD_{ss} values.



(**a**) (**b**)

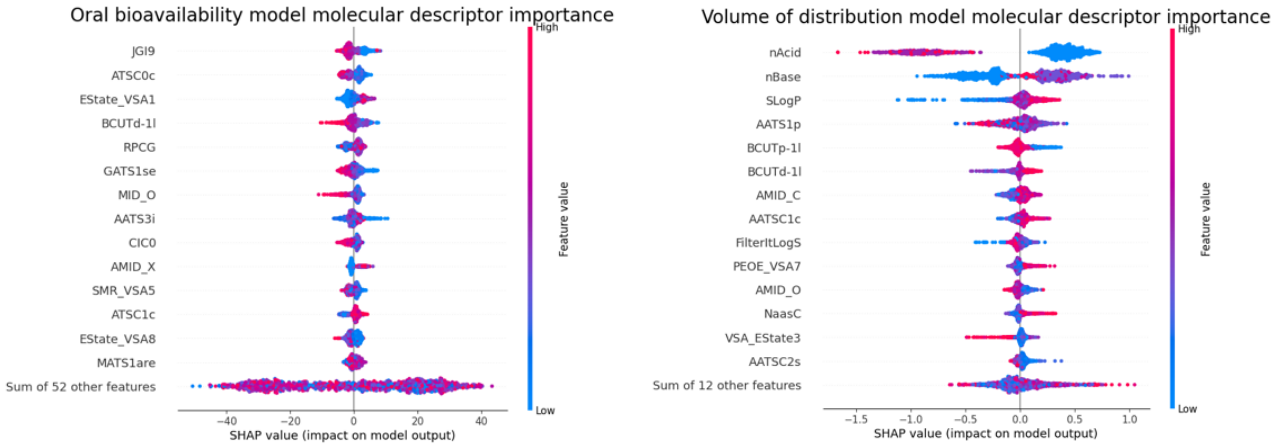
Figure 3. The effect of different definitions of applicability domains on coverage of validation set 1 and predictive performance (**a**) The 3-NN Tanimoto AD. The evolution of the GMFE (in red) on validation set 1 according to different Tanimoto thresholds ranging from 0 to 1. The evolution of the coverage of the validation set is plotted as blue bars. (**b**) The SARpy AD. The evolution of the GMFE performance (in red) on the validation set according to different Log ratio thresholds relative to the structural alert associated with the prediction. The evolution of the effective size retained in the validation set is plotted alongside in blue. The plot was made using the SARpy model with the 0.6–5 $L \cdot kg^{-1}$ threshold.

2.4. Molecular Descriptor Importance

The importance of the molecular descriptors in the best models was analyzed using the SHAP (SHapley Additive exPlanations) values with the SHAP Python package (version 0.44.0) [35]. The SHAP value for each molecular descriptor (in rows) indicates the degree to which a model's computed predictions change when the values of molecular descriptors vary. In Figure 4, all the SHAP values for the top 15 molecular descriptors are displayed in rows. The x-axis represents the SHAP values, while the y-axis depicts the molecular descriptors, ordered by importance from highest (at the top) to lowest (at the bottom). Each dot corresponds to a chemical and is color-coded according to the value of the corresponding molecular descriptor, ranging from high to low.

Among the top 15 most important molecular descriptors for the R-CatBoost regression oral bioavailability model (Figure 4a), we observe complex molecular descriptors that retain topological and electrostatic information, with the JGI9 (9-ordered mean topological charge), ATSC0c (centered Moreau–Broto autocorrelation of lag 0 weighted by Gasteiger charge), Estate_VSA1 (Labute's Approximate Surface Area EState indices and surface area), BCUTd-1I (first lowest eigenvalue of Burden matrix weighted by sigma electrons), and MID_O (molecular ID on O atoms) molecular descriptors being of most importance in the model.

These molecular descriptors are consistent with those identified in previous models developed for oral bioavailability. For instance, the model by Wei et al. [9] highlighted SsOH (an E-state molecular descriptor), ATS5i (a topological structure molecular descriptor), and TopoPSA(NO) as the most important. Similarly, the model by Ma et al. [16] identified additional topological structure descriptors, including TopoPSA and TopoPSA(NO), along with an E-state molecular descriptor (EState_VSA8) and the MID_O molecular descriptor, which is related to the identification and characterization of oxygen atoms in chemicals.



(**a**) (**b**)

Figure 4. (**a**) A summary plot obtained using the SHAP package. The plot shows the importance of the 15 most important molecular descriptors of the R-CatBoost (regression) oral bioavailability model and their effects on predictions. The plot depicts the relationship between a molecular descriptor's value and its impact on the prediction. For instance, high values of jGI9 (a topological charge molecular descriptor) are associated with a tendency to decrease oral bioavailability. (**b**) A summary plot obtained using the SHAP package for the R-RF (regression) VD_{ss} model. The plot shows the importance of the 15 most important molecular descriptors and their effect on the predictions. This plot depicts the relationship between a molecular descriptor value and its impact on the prediction. For example, high values of SLogP tend to increase VD_{ss}.

For VD_{ss}, Figure 4b depicts the molecular descriptor importance for the R-RF regression model. Among the top 15 molecular descriptors, we observe those that impact the model's prediction. We observe that low numbers of acidic groups increase VD_{ss} and low numbers of base groups decrease VD_{ss}. Another important molecular descriptor is the logarithm of the n-octanol–water partition coefficient (SLogP), an important factor in pharmacokinetics. These molecular descriptors were previously found to impact VD_{ss} [23].

The list of molecular descriptors, along with their Mordred molecular descriptions and examples of chemicals with high and low values, is provided in Supplementary Tables S10 and S11.

2.5. QSAR Mapping of EDCs as a Function of Key TK Properties

In order to characterize the TK profiles of chemicals regarded as endocrine disruptors (EDCs), we predicted key TK properties for 131 EDCs by applying three QSAR models: the two QSAR models for oral bioavailability and VD_{ss} described in this manuscript, using the SARpy AD, and an existing QSAR (from VEGA) model predicting the total body elimination half-life, for which we considered moderate and good experimental predictions to be inside the AD.

The oral bioavailability and volume-of-distribution prediction results for the targeted EDCs (categorized into 10 common chemical families) are shown in Figure 5, in addition to the total body elimination half-life prediction.

Among the studied chemical categories, perfluoro (alkyl/alkane) substances (PFASs) exhibited a long total body elimination half-life, suggesting prolonged retention in the body. However, these compounds typically had a low VD_{ss}, with the exception of PFASFs (Supplementary Figure S5), which demonstrated a moderate predicted VD_{ss}. Bisphenols, on the other hand, exhibited a moderate VD_{ss}, indicating a balanced distribution across tissues, and displayed medium oral bioavailability. These compounds were characterized by a relatively short elimination half-life, implying faster clearance from the body compared to PFASs.

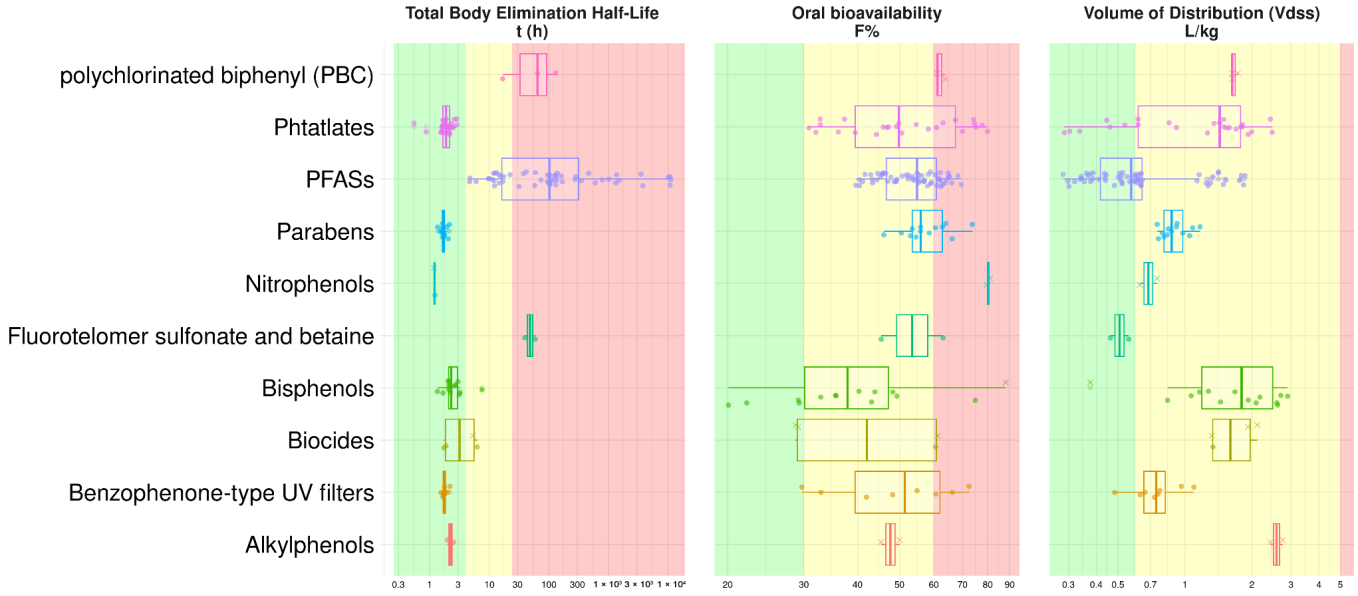


Figure 5. Boxplots of the predictions for oral bioavailability, VD_{ss}, and elimination half-life for a set of 131 EDCs categorized into 10 chemical categories. Perfluoroalkylcarboxylic acid (PFCA), perfluoroalkylsulfonic acid (PFSA), perfluoroalkylether acid (PFEA), perfluoroalkane sulfonamide derivative (PASF), perfluoroalkyl phosphonic acid (PFPiA), and polyfluoroalkyl phosphate diester (diPAP) categories were grouped as PFASs. Chemicals inside the SARpy AD are represented as circles, and chemicals outside as crosses. Background colors are set to green, orange, and red for, respectively, low, medium, and high values of VD_{ss} (thresholds: 0.6 $L{\cdot}kg^{-1}$ and 5 $L{\cdot}kg^{-1}$), oral bioavailability (thresholds: 30% and 60%), and elimination half-life (thresholds: 4 h and 24 h).

To assess the molecular descriptors driving model predictions, we visualized a heatmap of the mean standardized descriptor values for EDC compounds, grouped by chemical family, based on the major descriptors used in the oral bioavailability and VD_{ss} models (Figure 6). The results reveal that PBCs and nitrophenols exhibit notably low values of GATS1se and AATS3i, features associated with high predicted oral bioavailability in the model. These characteristics could contribute to the model assigning elevated oral bioavailability to compounds in these families. PFASs display low values of BCUTp-1l, AMID_C, and AATSC2s, combined with high nAcid and low nBase counts. The model predicted a low volume of distribution (VD_{ss}) for these chemical families, suggesting that these structural traits are key drivers of the pharmacokinetic behavior predicted for these EDCs.

Interestingly, only a few chemicals were inside the AD of the VD_{ss} and oral bioavailability models, with, respectively, 16 and 27 chemicals inside the 3-NN Tanimoto AD. For the elimination half-life, where a good prediction is considered to be inside the applicability domain, 24 chemicals were considered. Among them, Phthalates (DBP, DCHP, DEP, DMP, ...), Benzophenone-type UV filters, 4-n-Nonylphenol (Alkylphenols), and Benzylparaben emerged for the VD_{ss} model, and Phthalates (DBP, DCHP, DEP, DMP, ...), 4-n-Nonylphenol (Alkylphenols), Benzophenone-type UV filters, Parabens, and Bisphenols (BPE, BPF, BPS) for the oral bioavailability model.

With the SARpy AD, of 131 chemicals, 118 and 91 were inside the AD for VD_{ss} and oral bioavailability, respectively. The method also allowed us to identify structural patterns among the groups of chemicals that were linked to high or low values of VD_{ss} or oral bioavailability. For example, aromatic rings or two aromatic rings linked, which are found in bisphenols, PBC, and benzophenone-type UV filters, are associated with medium or high values of VD_{ss}. An aromatic ring linked to a carboxylic group, found in parabens and phthalates, is associated with low values of VD_{ss} (Figure 7a). Perfluoroalkyl groups, found

in PFASs, are associated with high oral bioavailability, and long carbon chains, found in parabens, are associated with low oral bioavailability (Figure 7b).

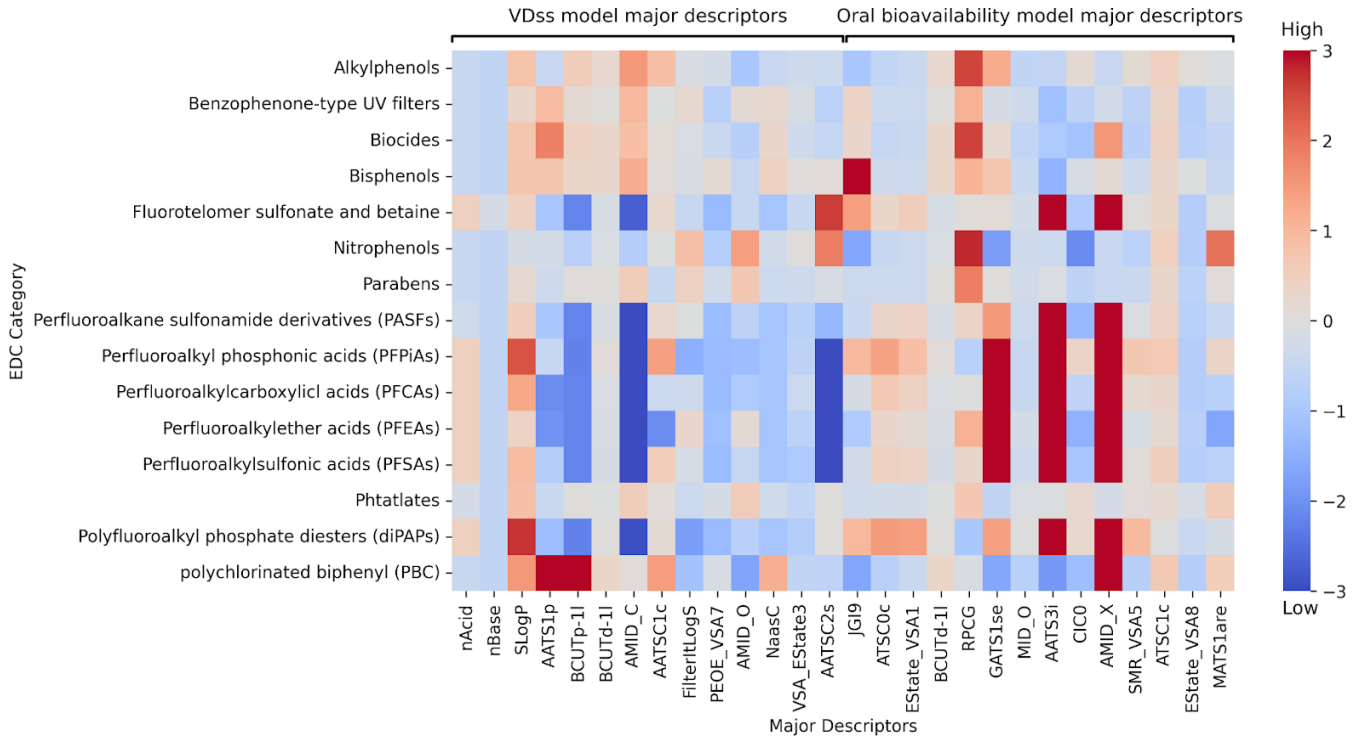
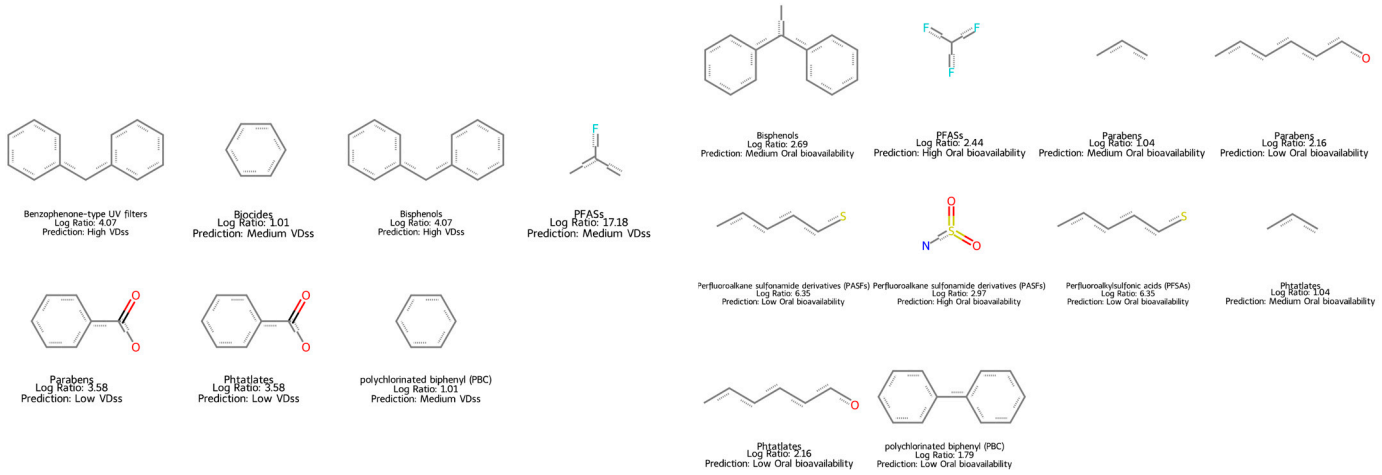


Figure 6. Heatmap of the mean standardized descriptor values for the EDC list, grouped by EDC families, based on the major descriptors of the oral bioavailability and VD_{ss} models. Standardization was performed using the mean and standard deviation computed from the training set.



(**a**) (**b**)

Figure 7. Structural fragment alerts identified by MC-SARpy in more than two chemicals across different EDC categories for the VD_{ss} (**a**) and oral bioavailability (**b**) models. The Log ratio and the predicted category associated with each structural alert are reported. PASFs, PFCAs, PFEAs, PFSAs, diPAPs, and PFPiAs were combined into the PFAS category when the structural fragment was identical.

In order to have an idea about the relevance of our predictive models for EDCs, we searched the literature for toxicokinetic (TK) profiles reported in humans. We found that TK profiles of bisphenols were assessed in piglets in the study by Gély et al. [36]. BPA and its alternatives exhibited low oral bioavailability, medium to high VD_{ss}, and a

short elimination half-life. Studies on humans have estimated the elimination half-life of deuterated BPA to be approximately 6.4 ± 2.0 h [37].

For benzophenone UV filters, the literature reports a short elimination half-life of around 4 h [38]. Per- and polyfluoroalkyl substances (PFASs) generally exhibit high oral bioavailability. For example, PFOA and EOF showed bioavailability values of 65–71% and 74–87%, respectively, in mouse studies [39]. In workers exposed to perfluoroalkyl surfactants, a low mean distribution volume of 0.08 $L \cdot kg^{-1}$ was reported [40]. Drew et al. [41] investigated the elimination half-lives of several PFASs, including PFOS, PFHpS, PFHxS, PFNA, and PFDA, reporting prolonged elimination half-lives of 74.1 ± 13.4 h, 45.7 ± 9.4 h, 9.3 ± 1.3 h, 12.3 ± 3.2 h, and 60.4 ± 10.4 h, respectively. Phthalates were found to have short elimination half-lives. For example, DEHP exhibited an elimination half-life of 4.3–6.6 h in humans [42]. Overall, these findings from the literature align with our TK QSAR models.

We also predicted TK properties for a set of 316 chemicals that are likely to disrupt ARs and ERs. Using dedicated thresholds for each TK property (VD_{ss}: 0.6 kg/L and 5 kg/L; oral bioavailability: 30% and 60%; elimination half-life: 4 h and 24 h), we set chemical attributes as low, medium, and high TK concern to highlight chemicals that are characterized by concerning TK profiles in terms of chemical risk. Among the 316 chemicals, 67.4% (213 chemicals) were inside the SARpy AD of the oral bioavailability QSAR, 70.9% (224 chemicals) were inside the SARpy AD of the VD_{ss} QSAR, and 94.3% (298 chemicals) were inside the ADI of the elimination half-life QSAR.

Among the 316 chemicals, 16.4% (52 chemicals) were predicted as having a TK risk, with at least one TK property classified as high.

Among them, Bisphenol AF was predicted to have high oral bioavailability, a medium half-life, and a medium VD_{ss}. This chemical poses a risk, as it is produced in large amounts—100 to 1000 tons, as stated by the European Chemical Agency (ECHA) [43]. Seven other chemicals were found to be registered in ECHA (4′,5′-Diiodofluorescein; 3,5-Dichloro-4-hydroxybenzophenone; 3-[1-[4-[2-(dimethylamino)ethoxy]phenyl]-2-phenylbut-1-enyl]phenol; 3′,6′-dihydroxyspiro [2-benzofuran-3,9′-xanthene]-1-one; 4′,5′-dibromo-3′,6′-dihydroxyspiro[isobenzofuran-1(3H),9′-[9H]xanthene]3-one; Bisphenol AF; mitotane; and 1-chloro-2-[2,2,2-trichloro-1-(4-chlorophenyl)ethyl]benzene). For example, the use of 4′,5′-Diiodofluorescein in cosmetic products was banned in Europe, 3-[1-[4-[2-(dimethylamino) ethoxy]phenyl]-2-phenylbut-1-enyl]phenol, and bisphenol AF were recognized as toxic to reproduction, and CLP describes 2,2,2,o,p′-pentachloroethylidenebisbenzene as fatal if inhaled and toxic if swallowed.

From the set of identified EDCs, three have a high TK risk, namely, (E,Z)-Tamoxifene, Clomiphene, and (E)-Toremifene, all having a high VD_{ss}, high oral bioavailability, and a medium body elimination half-life. These chemicals are known to be related to endocrine disruption: Clomiphene is a drug that increases the chance of pregnancy by facilitating ovulation [44], Tamoxifen is a drug used to treat hormone-positive breast cancer [45], and Toremifene is known to bind to estrogen receptors and act as a weak partial agonist and potent antagonist [46]. Overall, these results show the relevance of using our QSAR models to predict EDCs and TK properties to identify chemicals that are most likely to pose a risk.

3. Discussion

This study used a large dataset comprising over 1600 chemicals to develop a QSAR model for both oral bioavailability and VD_{ss} for regression, binary classification, and multiclass prediction.

Among similar studies considering oral bioavailability at a 50% threshold, Falcón-Cano et al. (2020) [10] employed a dataset of over 1400 compounds and achieved a BA of

0.78. Venkatraman (2021) [11] used 1800 chemicals and reported a BA of 0.71, and Wei et al. (2022) [9] achieved an accuracy of 0.79. Our model exhibited comparable results to these studies, with a BA of 0.77, corresponding to an accuracy of 0.77 on a different validation set. Recently, Ma et al. (2024) [16] reported an accuracy of 0.82 using the same 209-compound validation set as Falcón-Cano et al. [10].

The QSAR model described in this article can be regarded as more robust than those previously published. In particular, our model was trained and evaluated on a larger dataset, with twice as many chemicals in our validation set compared to those used by Falcón-Cano et al., Wei et al., and Ma et al. (405 in our study vs. 209) [10,11,16].

For VD_{ss} model development in related work with similar numbers of substances that used the GMFE metric to evaluate their models, Lombardo et al. (2021) [17,18], who had fewer compounds in the training set and used the same validation set of 34 compounds, reported a GMFE of 1.70, while our regression model exhibited a GMFE of 1.81 (Supplementary Table S4a).

Our results are therefore comparable to those of previously published models and, as discussed for bioavailability, can be considered more robust given the larger training set size. In addition, we were able to model and compare the development of regression, binary classification, and multiclass classification models for this endpoint. The development of an applicability domain to determine the limit of our QSAR models and the application of SARpy resulted in some structural fragment alerts on EDCs that were linked to high or low values of VD_{ss} or oral bioavailability.

The development of our QSAR models followed the OECD QSAR validation principles. In line with Principle 1, the models have defined endpoints for oral bioavailability and VD_{ss}. Principle 2 is addressed with an unambiguous algorithm (scripts available as Supporting Material together with model reporting formats, QMRF), defined methods, and the retained models—R-CatBoost, BC-CatBoost, and MC-CatBoost—for the prediction of oral bioavailability, corresponding to continuous value prediction, binary classification prediction, and multiclass prediction, respectively. For VD_{ss}, the retained models are R-RF, BC-Chemprop, and MC-Chemprop, corresponding to continuous value prediction, binary classification prediction, and multiclass prediction, respectively. Principle 3 is addressed by an applicability domain, defined using two approaches: a structural alert approach using SARpy and an analog-based approach using a three-nearest neighbor method.

The models follow Principle 4 by ensuring appropriate measures of goodness of fit, robustness, and predictivity. This was demonstrated by strong performance on the training set (seen data) and the external validation set (unseen data), as well as through 50 iterations of five-fold cross-validation.

Principle 5, which concerns the definition of a mechanistic interpretation, is explored through model molecular descriptor importance and the applicability domain. The applicability domain both explores the nearest neighbors and allows identification of the most important structural fragments contributing to predictions with the SARpy models. Both approaches can be used together to define the AD, each providing deeper insight into the prediction. However, we recommend using the SARpy LR AD approach, as it offers a clearer understanding of the structural fragments responsible for the activity. Another definition of the applicability domain was explored using Insubria plots, relying on the leverage approach from the hat matrix [47]. Similarly to 3-NN, we can observe that around 95% of the compounds fall inside the applicability domain for oral bioavailability and VD_{ss} (Supplementary Figure S6).

4. Materials and Methods

4.1. Data

4.1.1. Oral Bioavailability Data Source

Data on human oral bioavailability were collected from multiple sources, including OCHEM [48], CHEMBL [49], and articles by Min Wei et.al [9], Falcón-Cano et al. [10], Varma et al. [50], and Tian et al. [13]. In total, 1712 chemicals and associated oral bioavailability data were retrieved and curated. Special attention was paid to the presence of qualifiers (greater or less than a certain threshold) for oral bioavailability in order to properly consider this information with respect to the categorization thresholds adopted during the discretization of continuous values.

4.1.2. Volume of Distribution Data Source

Data on human VD_{ss} values were collected from CHEMBL [49], the article by Lombardo et al. [18], and the article by Liu et al. [22]. In total, 1591 chemicals and associated VD_{ss} data were retrieved and curated.

4.1.3. Preprocessing Standardization

All the chemicals were mapped to their PubChem Compound ID (CID) in order to harmonize chemical structures (SMILES) that are standardized according to the PubChem protocol [51] (i.e., normalization of the representation, implicit hydrogens, atom valence, tautomeric form representation, etc.). The PubChem CID was retrieved according to the available SMILES, CAS RN, name, and InChI available from the source database. In cases of chemicals with ions, the largest fragment was considered. This standardization allowed us to identify duplicate chemicals for which we computed the mean F% and mean VD_{ss} values. Duplicate chemicals with a difference greater than 20 for the standard deviation of F%, as well as those with F% values exceeding 100 or falling below 0, were excluded from the dataset.

4.2. Dataset Preparation for Modeling

4.2.1. Oral Bioavailability Data Preparation

The datasets were randomly split into training and validation sets. To construct the validation set, chemicals were sorted based on their F%, and we selected every fourth chemical to populate this set. This choice ensured representative inclusion across the range of F% values. This approach resulted in ~25% of chemicals (405 chemicals) being selected for the validation set.

The remaining 75% of chemicals of the training set with known F% values were retained for regression modeling (1213 chemicals). Some chemicals in the literature did not have F% values but only qualitative information about low or high F% with respect to different thresholds (50%, 30%, and 60%). For the implementation of QSAR classifiers, chemicals with known threshold values were considered. This encompassed the design of a training set for binary classification (50% threshold, distinguishing low and high classes with 1307 chemicals) and another training set for multiclass classification (30–60% thresholds, defining low, medium, and high classes with 1244 chemicals). The 50% thresholds (single-threshold classification models) were used as described in the literature [9,10], and the 30–60% thresholds (double-threshold multiclass models) were used considering the thresholds adopted by some CRO experts in this domain (personal communication). No chemicals from the respective training set were included in the validation set.

4.2.2. Volume of Distribution Data Preparation

The datasets for volume of distribution were randomly split into training and validation sets, similarly to the oral bioavailability data. This resulted in ~25% of chemicals being selected for the validation set (390 chemicals). A second validation set was developed using a list of 34 chemicals extracted from the article by Lombardo et al. (2016) [17]. This dataset, which did not contain chemicals belonging to sets used with previously published models, was used for model comparison. No chemicals from the two validation sets were included in the training set. Here, the natural logarithm ($\ln(VD_{ss})$) was used to facilitate QSAR model development [20].

The classification model was developed on the training set (1167 chemicals), with a threshold of 1 $L \cdot kg^{-1}$ corresponding to a chemical extensively (90%) distributed in tissues [52]. For multiclass classification, we considered thresholds of 0.6 $L \cdot kg^{-1}$ and 5 $L \cdot kg^{-1}$ [7].

4.3. Molecular Descriptors

We computed Mordred molecular descriptors [53] covering a wide range of structural and physico-chemical properties of chemicals. A total of 1826 molecular descriptors were computed. Subsequently, columns containing "NA" values, molecular descriptors with zero variance, and those exhibiting absolute pair correlations exceeding 0.97 were excluded from the dataset to reduce redundant information. This process yielded 560 molecular descriptors for the oral bioavailability regression dataset, 507 for the binary classification dataset, 560 for the multiclass classification dataset, and 500 for the volume-of-distribution datasets.

4.4. Selection of Molecular Descriptors

In order to reduce the number of molecular descriptors (i.e., the independent variables of the models) and increase the parsimony and interpretability of the models, the VSURF algorithm [54] was applied to select and retain only the most informative molecular descriptors. The R package VSURF (version 1.2.0) allows identification of the most informative molecular descriptors using random forest importance scores based on permutation and using a stepwise forward strategy that selects the variables of the most accurate models. Molecular descriptor selection was performed exclusively on the training sets to avoid data leakage. VSURF identifies two sets of molecular descriptors: interpretation and prediction levels. We selected the interpretation set, as it contains the most molecular descriptors. We measured the Topliss ratio as the number of training chemicals per molecular descriptor, in accordance with OECD guidelines, which recommend a rule-of-thumb ratio greater than 5 [55].

4.5. Machine Learning Algorithms

We considered a variety of machine learning models, such as CatBoost [56], XGBoost [57], random forest (RF) [58], and Chemprop [59]. The SARpy [60,61] method, a structure–activity relationship (SAR) method, was also tested for prediction to enhance the mechanistic interpretability of QSAR models. A description of each machine learning approach is available in Supplementary Materials (References [62–64] are cited in the Supplementary Materials). CatBoost, XGBoost, chemprop, and RF were applied to regression, binary classification, and multiclass classification. SARpy was used for binary and multiclass classification.

4.6. Protocol

The tuning of algorithm hyperparameters was optimized using the training set without using any data from the validation sets.

For CATBoost, XGBoost, and RF, 5-fold cross-validation was carried out for hyperparameter optimization using a grid search (Supplementary Table S1).

After this step, the models were subjected to 50 iterations of 5-fold cross-validation using the training set in order to assess algorithm robustness and identify the best-performing algorithm. Specifically, the training set was partitioned into five non-overlapping subsets 50 times. In each iteration, four subsets were combined to train the model, while the remaining subset was used to evaluate model performance in cross-validation.

The predictive performance of the "unseen" chemicals in the validation datasets was then assessed according to commonly used statistical indicators, i.e., sensitivity (SE), specificity (SP), and balanced accuracy (BA) for the binary classification; class-specific SE, SP, and BA, macro/micro-SE, macro/micro-SP, and macro/micro-Ba for multiclass classification; Root Mean Squared Error (RMSE), mean absolute error (MAE), Q^2_{F3}, and R^2 for oral bioavailability regression models; and geometric mean fold error (GMFE) for VD_{ss} models. The regression model with the best performance on the external validation set was also evaluated for its performance in binary and multiclass classification tasks (Supplementary Figure S2).

4.7. Predictive Performance Metrics

The models were evaluated externally and internally using their respective training and validation sets, whose data were not used to calibrate or optimize the models.

The performance metrics adopted for classification are defined as follows:

$$\text{Sensitivity (SE)} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Specificity (SP)} = \frac{TN}{TN + FP} \tag{2}$$

$$\text{Balanced accuracy (BA)} = \frac{Sensitivity + Specificity}{2} \tag{3}$$

where *TP* stands for True Positive, *TN* for True Negative, *FN* for False Negative, and *FP* for False Positive. For multiclass prediction, these same metrics were computed for each category at a time by considering the category under scrutiny as active.

For multiclass classification models, we evaluated performance using metrics calculated individually for each class by treating it as the positive class, providing class-specific SE, SP, and BA. Additionally, we computed micro-averaged metrics, which aggregate all instances to give equal weight to each sample, and macro-averaged metrics, which average performance across classes equally, ensuring balance between majority and minority classes. This combination offers both detailed class-level insights and an overall performance assessment.

For the oral bioavailability regression models, the performance metrics were defined as follows [65]:

$$RMSE = \sqrt{\frac{\sum(yi - \hat{y}i)^2}{n}} \tag{4}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}\left(yi - \underline{y}\right)^2} \tag{5}$$

$$Q^2_{F3}(y,\hat{y}) = 1 - \frac{\left(\sum_{i=1}^{n(OUT)}(y_i - \hat{y}_i)^2\right)/n_{OUT}}{\left(\sum_{i=1}^{n(TR)}(y_i - \hat{y}_{TR})^2\right)/n_{TR}} \quad (6)$$

$$MAE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |yi - \hat{y}i| \quad (7)$$

For the modeling of VD_{ss}, we assessed the performance of the regression models using the geometric mean fold error (GMFE) [17,18].

The GMFE was computed as follows:

$$GMFE(y,\hat{y}) = 10^{(\sum_{i=1}^{n} |Log_{10}(\frac{\hat{y}}{yi})|)/n} \quad (8)$$

where $\hat{y}i$ is the predicted value of the i-th sample, and yi is its corresponding experimental value; $\underline{y}$ is the mean of the predicted values. n_{TR} and n_{OUT} are the number of training and validation chemicals, respectively; $\underline{y}_{TR}$ is the average value of the training set experimental responses; and $\underline{y}_{OUT}$ is the average value of experimental responses of external validation chemical values.

The VD_{ss} regression models were trained using logarithmic values, and the linear predicted values were subsequently adopted for the GMFE formula. GMFE is a standard metric for evaluating PK models when the model is trained with values in the logarithmic space and the predicted values are recovered through $y = e^{y(log)}$, where y(log) is the predicted value in the logarithmic space [66]. GMFE values around and below 2 are generally considered indicators of acceptable precision for pharmacokinetic parameters [34].

The predictive performance of the regression models was also evaluated in terms of classification. For this evaluation, the validation set was predicted with the best model and then classified according to the corresponding thresholds of oral bioavailability and VD_{ss}.

4.8. Definition of the Applicability Domain

4.8.1. K-Nearest Neighbors

In QSAR modeling, the applicability domain refers to the precision of the prediction computed by a model within a given chemical space, thereby providing information about the expected level of reliability of predictions computed for unseen chemicals included in the molecular descriptor space defined by the applicability domain (AD). The applicability domain helps prevent model misuse and enhances the trustworthiness of predictions. It is established by considering the training set and serves as a guideline for determining which chemicals the model can assess with a given reliability [67].

Various methods exist for defining the applicability domain in QSAR. In this study, a distance-based approach was chosen (3-NN Tanimoto AD), evaluating the similarity between a query chemical and those in the training set. This similarity was measured using the Tanimoto score, which measures chemical similarity using chemicals encoded as fingerprints; here, we used Morgan fingerprints [62]. For each query chemical in the validation set, we calculated the average Tanimoto score of the three most similar compounds from the training set. The models' predictive performance was then characterized by analyzing the precision of predictions across different threshold values for the Tanimoto score.

Establishing a useful threshold to determine whether a chemical falls within the applicability domain requires balancing precision and coverage of the chemical space. Various methods exist for setting this threshold, and in this study, we defined it as Dc = <y> − Zσ. Here, ⟨y⟩ represents the average Tanimoto score of the three closest training set neighbors for each chemical, while σ\sigmaσ denotes the standard deviation of these scores. The parameter Z controls the significance level, with a default value of

0.5. Unlike the original formula, which uses Euclidean distance (where 0 signifies identical chemicals), our approach subtracts Zσ from ⟨y⟩, adapting it to the Tanimoto score scale (ranging from 0 to 1, where 1 indicates identical chemicals) [68,69].

4.8.2. SARpy

We also investigated the possibility of applying SARpy to define the applicability domain (SARpy AD). For this purpose, we took into account the likelihood ratio (LR) for each structural alert (SA) associated with each predicted chemical to gauge its precision in correctly predicting the chemical. A chemical was considered to be within the applicability domain if the LR of the structural alert responsible for the predicted compound exceeded a specific threshold. The predictive performance of the models was then compared by evaluating their effectiveness across various LR threshold values.

4.9. Mapping of EDCs

We mapped the TK space of EDCs on two lists of chemicals, the first one containing a selection of 131 endocrine-disrupting chemicals reported in the literature [70]. These EDCs can be found in everyday life in different products, including additives, food packaging, food and beverage containers or cans, cosmetics, cookware, toys, hygiene and cleaning products, etc. [71–73]. Many of these chemicals can be found at detectable levels in the urine and blood of children and adults [74–76].

The second set, consisting of 55,450 chemicals to which humans are potentially exposed, forms a list of toxicological and environmental chemicals of interest [77]. From this set, we applied QSAR models for estrogen binding [78] and androgen binding [79] and selected the chemicals most likely to perturb the considered receptors by considering only QSAR predictions characterized by an applicability domain index greater than >0.8. This strict requirement resulted in a subset of 316 chemicals being retained for screening.

Oral bioavailability and VD_{ss} were computed by the QSAR models described herein. In addition, the elimination half-life was computed using the freely available dedicated VEGA QSAR [78] model, and the TK properties of different groups of EDCs were compared.

All the models and code for the prediction of oral bioavailability and volume of distribution at steady state are available at https://github.com/guillaumeolt/QSAR_TK (accessed on 16 September 2025).

5. Conclusions

In this study, we developed QSAR models to predict oral bioavailability and VD_{ss} using state-of-the-art machine learning approaches and following OECD guidelines. Leveraging current databases for these pharmacokinetic endpoints, we developed regression, binary classification, and multiclass classification models and systematically evaluated their performance across various scenarios using relevant metrics. We applied a 3-NN applicability domain approach and highlighted the importance of using SAR methods to define applicability domains. Furthermore, we integrated these QSAR models with a complementary elimination half-life model and applied them to a curated list of endocrine disruptors and a list of toxicological and environmental chemicals of interest. This combined approach identified critical categories and chemicals of concern, providing valuable insights for the prioritization and regulatory evaluation of endocrine-disrupting chemicals.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jox15050166/s1, Table S1: Optimized parameters for the machine learning models, Table S2a: Performance table of the regression model of oral bioavailability, Table S2b: Cross validation performance table of the regression model of oral bioavailability, Table S3a: Performance table of the classification model of oral bioavailability, Table S3b: Cross

validation performance table of the classification model of oral bioavailability, Table S4a: Performance table of the multiclass classification model of oral bioavailability, Table S4b: Cross validation performance table of the multiclass classification model of oral bioavailability, Table S5a: Performance table of the regression model of VD_{ss}, Table S5b: Cross validation performance table of the regression model of VD_{ss}, Table S6a. Performance table of the classification model of VD_{ss}, Table S6b: Cross validation performance table of the classification model of VD_{ss}, Table S7a: Performance table of the multiclass classification model of VD_{ss}, Table S7b: Cross validation performance table of the multiclass classification model of VD_{ss}, Table S8: Table of the best structural alerts found for the multiclass prediction of oral bioavailability, Table S9: Table of the best structural alerts found for the multiclass prediction of VD_{ss}, Figure S1a: UMAP representation of the chemical space in a 2D map projection for the oral bioavailability dataset. Each point represents a chemical, and its color encodes the corresponding F% value, ranging from red for low oral bioavailability to blue for high oral bioavailability; 32% of chemicals have F% values below 30 (F%), 45% above 60 (F%), and 23% between 30 (F%) and 60 (F%), Figure S1b: UMAP representation of the chemical space in a 2D map projection for the VD_{ss} dataset. Points are colored considering the Log-transformed VD_{ss} values from red to blue for low to high VD_{ss}; 37% of chemicals have VD_{ss} values below 0.6 $L{\cdot}kg^{-1}$, 16% above 5 $L{\cdot}kg^{-1}$, 47% between 0.6 $L{\cdot}kg^{-1}$, and 5 $L{\cdot}kg^{-1}$, Figure S2: General protocol applied for the development and evaluation of predictive models for the prediction of oral bioavailability and VD_{ss}, Figure S3: Predicted vs. true oral bioavailability values on the 405 chemicals of the validation set. Red and blue dashed lines correspond to a 10% and 20% error, respectively, Figure S4: Predicted vs. true VD_{ss} values on the 405 chemicals of the validation set. Red and blue dashed lines correspond to a 2-fold and 3-fold error, respectively, Figure S5: Boxplot of the predictions for oral bioavailability, VD_{ss} and elimination half-life for a set of EDC categorized by chemical category, Figure S6: Plot of the predicted values against the hat values for the (a) oral bioavailability R-CatBoost model and the (b) VD_{ss} R-RF models, Table S10: Table of the best molecular descriptors for the best model of regression of oral bioavailability, Table S11: Table of the best molecular descriptors for the best model of regression of VD_{ss}, Table S12: Table of models' regression prediction on a set of toxicological and environmental chemical list of interest, Table S13: Table of models' regression prediction on a list of endocrine disruptors, Table S14: Table of models' regression prediction with molecular descriptors on a set of toxicological and environmental chemical list of interest, Table S15: Dataset for oral bioavailability, Table S16: Dataset for VD_{ss}, File S1: QMRF VD_{ss}, File S2: QMRF Oral bioavailability. References [63,64,80–82] are cited in Supplementary Materials.

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

Funding: The ED-SCREEN project (ANSES-21-EST-131) was funded by the French National Research Program for Environmental and Occupational Health and supervised by the French Agency for Food, Environmental, and Occupational Health and Safety (Anses).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Acknowledgments: We would like to thank Pierre-André Billat for personal communication on the thresholds to consider for oral bioavailability and volume of distribution.

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

Abbreviations

The following abbreviations are used in this manuscript:

R	Regression
BC	Binary Classification
MC	Multiclass Classification
CV	Cross Validation
TK	Toxicokinetics
PK	Pharmacokinetics
GMFE	Geometric Mean Fold Error
QSAR	Quantitative Structure–Activity Relationship
VD	Volume of Distribution
VD_{ss}	Volume of Distribution at Steady State
EDC	Endocrine-Disrupting Chemicals
SE	Sensitivity
SP	Specificity
BA	Balanced Accuracy
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
3-NN	Three Nearest Neighbors
AD	Applicability Domain
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction

References

1. Shanmugam, P.S.T.; Sampath, T.; Jagadeeswaran, I.; Bhalerao, V.P.; Thamizharasan, S.; Krithaksha, V.; Saha, J. Toxicokinetics. In *Biocompatibility Protocols for Medical Devices and Materials*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 175–186, ISBN 978-0-323-91952-4.
2. Coecke, S.; Pelkonen, O.; Leite, S.B.; Bernauer, U.; Bessems, J.G.; Bois, F.Y.; Gundert-Remy, U.; Loizou, G.; Testai, E.; Zaldívar, J.-M. Toxicokinetics as a Key to the Integrated Toxicity Risk Assessment Based Primarily on Non-Animal Approaches. *Toxicol. Vitr.* **2013**, *27*, 1570–1577. [CrossRef]
3. Gundert-Remy, U.; Sonich-Mullin, C. The Use of Toxicokinetic and Toxicodynamic Data in Risk Assessment: An International Perspective. *Sci. Total Environ.* **2002**, *288*, 3–11. [CrossRef] [PubMed]
4. Roberts, D.M.; Buckley, N.A. Pharmacokinetic Considerations in Clinical Toxicology: Clinical Applications. *Clin. Pharmacokinet.* **2007**, *46*, 897–939. [CrossRef] [PubMed]
5. Price, G.; Patel, D.A. Drug Bioavailability. [Updated 30 July 2023]. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2025. Available online: https://www-ncbi-nlm-nih-gov.ezproxy.u-paris.fr/books/NBK557852/ (accessed on 16 September 2025).
6. Li, W.; Picard, F. Toxicokinetics in Preclinical Drug Development of Small-molecule New Chemical Entities. *Biomed. Chromatogr.* **2023**, *37*, e5553. [CrossRef]
7. Smith, D.A.; Beaumont, K.; Maurer, T.S.; Di, L. Volume of Distribution in Drug Design: Miniperspective. *J. Med. Chem.* **2015**, *58*, 5691–5698. [CrossRef]
8. Mansoor, A.; Mahabadi, N. Volume of Distribution. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2025.
9. Wei, M.; Zhang, X.; Pan, X.; Wang, B.; Ji, C.; Qi, Y.; Zhang, J.Z.H. HobPre: Accurate Prediction of Human Oral Bioavailability for Small Molecules. *J. Cheminform.* **2022**, *14*, 1. [CrossRef]
10. Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M.Á. ADME Prediction with KNIME: Development and Validation of a Publicly Available Workflow for the Prediction of Human Oral Bioavailability. *J. Chem. Inf. Model.* **2020**, *60*, 2660–2667. [CrossRef]
11. Venkatraman, V. FP-ADMET: A Compendium of Fingerprint-Based ADMET Prediction Models. *J. Cheminform.* **2021**, *13*, 75. [CrossRef]
12. Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; et al. ADMETlab 2.0: An Integrated Online Platform for Accurate and Comprehensive Predictions of ADMET Properties. *Nucleic Acids Res.* **2021**, *49*, W5–W14. [CrossRef] [PubMed]
13. Tian, S.; Li, Y.; Wang, J.; Zhang, J.; Hou, T. ADME Evaluation in Drug Discovery. 9. Prediction of Oral Bioavailability in Humans Based on Molecular Properties and Structural Fingerprints. *Mol. Pharm.* **2011**, *8*, 841–851. [CrossRef]

14. Kim, M.T.; Sedykh, A.; Chakravarti, S.K.; Saiakhov, R.D.; Zhu, H. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches. *Pharm. Res.* **2014**, *31*, 1002–1014. [CrossRef]
15. Musther, H.; Olivares-Morales, A.; Hatley, O.J.D.; Liu, B.; Rostami Hodjegan, A. Animal versus Human Oral Drug Bioavailability: Do They Correlate? *Eur. J. Pharm. Sci.* **2014**, *57*, 280–291. [CrossRef]
16. Ma, L.; Yan, Y.; Dai, S.; Shao, D.; Yi, S.; Wang, J.; Li, J.; Yan, J. Research on Prediction of Human Oral Bioavailability of Drugs Based on Improved Deep Forest. *J. Mol. Graph. Model.* **2024**, *133*, 108851. [CrossRef]
17. Lombardo, F.; Jing, Y. In Silico Prediction of Volume of Distribution in Humans. Extensive Data Set and the Exploration of Linear and Nonlinear Methods Coupled with Molecular Interaction Fields Descriptors. *J. Chem. Inf. Model.* **2016**, *56*, 2042–2052. [CrossRef]
18. Lombardo, F.; Bentzien, J.; Berellini, G.; Muegge, I. In Silico Models of Human PK Parameters. Prediction of Volume of Distribution Using an Extensive Data Set and a Reduced Number of Parameters. *J. Pharm. Sci.* **2021**, *110*, 500–509. [CrossRef]
19. Gombar, V.K.; Hall, S.D. Quantitative Structure–Activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution. *J. Chem. Inf. Model.* **2013**, *53*, 948–957. [CrossRef]
20. Fagerholm, U.; Hellberg, S.; Alvarsson, J.; Arvidsson McShane, S.; Spjuth, O. In Silico Prediction of Volume of Distribution of Drugs in Man Using Conformal Prediction Performs on Par with Animal Data-Based Models. *Xenobiotica* **2021**, *51*, 1366–1371. [CrossRef] [PubMed]
21. Simeon, S.; Montanari, D.; Gleeson, M.P. Investigation of Factors Affecting the Performance of in Silico Volume Distribution QSAR Models for Human, Rat, Mouse, Dog & Monkey. *Mol. Inform.* **2019**, *38*, 1900059. [CrossRef]
22. Liu, W.; Luo, C.; Wang, H.; Meng, F. A Benchmarking Dataset with 2440 Organic Molecules for Volume Distribution at Steady State. *arXiv* **2022**, arXiv:2211.05661. [CrossRef]
23. Skakkebæk, N.E.; Lindahl-Jacobsen, R.; Levine, H.; Andersson, A.-M.; Jørgensen, N.; Main, K.M.; Lidegaard, Ø.; Priskorn, L.; Holmboe, S.A.; Bräuner, E.V.; et al. Environmental Factors in Declining Human Fertility. *Nat. Rev. Endocrinol.* **2022**, *18*, 139–157. [CrossRef] [PubMed]
24. Soto, A.M.; Sonnenschein, C. Endocrine Disruptors: DDT, Endocrine Disruption and Breast Cancer. *Nat. Rev. Endocrinol.* **2015**, *11*, 507–508. [CrossRef]
25. Heindel, J.J.; Newbold, R.; Schug, T.T. Endocrine Disruptors and Obesity. *Nat. Rev. Endocrinol.* **2015**, *11*, 653–661. [CrossRef]
26. Macedo, S.; Teixeira, E.; Gaspar, T.B.; Boaventura, P.; Soares, M.A.; Miranda-Alves, L.; Soares, P. Endocrine-Disrupting Chemicals and Endocrine Neoplasia: A Forty-Year Systematic Review. *Environ. Res.* **2023**, *218*, 114869. [CrossRef]
27. Ahn, C.; Jeung, E.-B. Endocrine-Disrupting Chemicals and Disease Endpoints. *Int. J. Mol. Sci.* **2023**, *24*, 5342. [CrossRef]
28. Calsolaro, V.; Pasqualetti, G.; Niccolai, F.; Caraccio, N.; Monzani, F. Thyroid Disrupting Chemicals. *Int. J. Mol. Sci.* **2017**, *18*, 2583. [CrossRef]
29. Goss, K.-U.; Brown, T.N.; Endo, S. Elimination Half-Life as a Metric for the Bioaccumulation Potential of Chemicals in Aquatic and Terrestrial Food Chains. *Environ. Toxicol. Chem.* **2013**, *32*, 1663–1671. [CrossRef] [PubMed]
30. Hallare, J.; Gerriets, V. Half Life. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2025.
31. Aungst, B.J. Optimizing Oral Bioavailability in Drug Discovery: An Overview of Design and Testing Strategies and Formulation Options. *J. Pharm. Sci.* **2017**, *106*, 921–929. [CrossRef] [PubMed]
32. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.
33. Wang, J.; Krudy, G.; Xie, X.-Q.; Wu, C.; Holland, G. Genetic Algorithm-Optimized QSPR Models for Bioavailability, Protein Binding, and Urinary Excretion. *J. Chem. Inf. Model.* **2006**, *46*, 2674–2683. [CrossRef]
34. Fendt, R.; Hofmann, U.; Schneider, A.R.P.; Schaeffeler, E.; Burghaus, R.; Yilmaz, A.; Blank, L.M.; Kerb, R.; Lippert, J.; Schlender, J.; et al. Data-driven Personalization of a Physiologically Based Pharmacokinetic Model for Caffeine: A Systematic Assessment. *CPT Pharmacomet. Syst. Pharmacol.* **2021**, *10*, 782–793. [CrossRef]
35. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
36. Gély, C.A.; Lacroix, M.Z.; Roques, B.B.; Toutain, P.-L.; Gayrard, V.; Picard-Hagen, N. Comparison of Toxicokinetic Properties of Eleven Analogues of Bisphenol A in Pig after Intravenous and Oral Administrations. *Environ. Int.* **2023**, *171*, 107722. [CrossRef]
37. Thayer, K.A.; Doerge, D.R.; Hunt, D.; Schurman, S.H.; Twaddle, N.C.; Churchwell, M.I.; Garantziotis, S.; Kissling, G.E.; Easterling, M.R.; Bucher, J.R.; et al. Pharmacokinetics of Bisphenol A in Humans Following a Single Oral Administration. *Environ. Int.* **2015**, *83*, 107–115. [CrossRef]
38. Stoeckelhuber, M.; Scherer, M.; Peschel, O.; Leibold, E.; Bracher, F.; Scherer, G.; Pluym, N. Human Metabolism and Urinary Excretion Kinetics of the UV Filter Uvinul A Plus® after a Single Oral or Dermal Dosage. *Int. J. Hyg. Environ. Health* **2020**, *227*, 113509. [CrossRef]

39. Gustafsson, Å.; Wang, B.; Gerde, P.; Bergman, Å.; Yeung, L.W.Y. Bioavailability of Inhaled or Ingested PFOA Adsorbed to House Dust. *Environ. Sci. Pollut. Res.* **2022**, *29*, 78698–78710. [CrossRef]
40. Fustinoni, S.; Mercadante, R.; Lainati, G.; Cafagna, S.; Consonni, D. Kinetics of Excretion of the Perfluoroalkyl Surfactant cC6O4 in Humans. *Toxics* **2023**, *11*, 284. [CrossRef]
41. Drew, R.; Hagen, T.G.; Champness, D.; Sellier, A. Half-Lives of Several Polyfluoroalkyl Substances (PFAS) in Cattle Serum and Tissues. *Food Addit. Contam. Part A* **2022**, *39*, 320–340. [CrossRef]
42. Kessler, W.; Numtip, W.; Völkel, W.; Seckin, E.; Csanády, G.A.; Pütz, C.; Klein, D.; Fromme, H.; Filser, J.G. Kinetics of Di(2-Ethylhexyl) Phthalate (DEHP) and Mono(2-Ethylhexyl) Phthalate in Blood and of DEHP Metabolites in Urine of Male Volunteers after Single Ingestion of Ring-Deuterated DEHP. *Toxicol. Appl. Pharmacol.* **2012**, *264*, 284–291. [CrossRef] [PubMed]
43. ECHA European Chemicals Agency. REACH—Registration, Evaluation, Authorisation and Restriction of Chemicals Regulation. 2025. Available online: https://echa.europa.eu/web/guest/information-on-chemicals/registered-substances (accessed on 11 March 2025).
44. Sovino, H.; Sir-Petermann, T.; Devoto, L. Clomiphene Citrate and Ovulation Induction. *Reprod. Biomed. Online* **2002**, *4*, 303–310. [CrossRef] [PubMed]
45. Cersosimo, R.J. Tamoxifen for Prevention of Breast Cancer. *Ann. Pharmacother.* **2003**, *37*, 268–273. [CrossRef] [PubMed]
46. Wiseman, L.R.; Goa, K.L. Toremifene: A Review of Its Pharmacological Properties and Clinical Efficacy in the Management of Advanced Breast Cancer. *Drugs* **1997**, *54*, 141–160. [CrossRef]
47. Gramatica, P.; Cassani, S.; Roy, P.P.; Kovarich, S.; Yap, C.W.; Papa, E. QSAR Modeling Is Not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)Triazoles on Algae. *Mol. Inform.* **2012**, *31*, 817–835. [CrossRef]
48. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V.V.; Tanchuk, V.Y.; et al. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided Mol. Des.* **2011**, *25*, 533–554. [CrossRef] [PubMed]
49. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef]
50. Varma, M.V.S.; Obach, R.S.; Rotter, C.; Miller, H.R.; Chang, G.; Steyn, S.J.; El-Kattan, A.; Troutman, M.D. Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *J. Med. Chem.* **2010**, *53*, 1098–1108. [CrossRef]
51. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51*, D1373–D1380. [CrossRef]
52. Toutain, P.L.; Bousquet-Mélou, A. Volumes of Distribution. *J. Vet. Pharmacol. Ther.* **2004**, *27*, 441–453. [CrossRef]
53. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10*, 4. [CrossRef]
54. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. VSURF: An R Package for Variable Selection Using Random Forests. *R J.* **2015**, *7*, 19. [CrossRef]
55. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; OECD Series on Testing and Assessment; OECD: Paris, France, 2014; ISBN 978-92-64-08544-2.
56. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *arXiv* **2019**, arXiv:1706.09516. [CrossRef]
57. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA; pp. 785–794.
58. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
59. Heid, E.; Greenman, K.P.; Chung, Y.; Li, S.-C.; Graff, D.E.; Vermeire, F.H.; Wu, H.; Green, W.H.; McGill, C.J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 9–17. [CrossRef]
60. Ferrari, T.; Gini, G.; Golbamaki Bakhtyari, N.; Benfenati, E. Mining Toxicity Structural Alerts from SMILES: A New Way to Derive Structure Activity Relationships. In Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Paris, France, 11–15 April 2011; pp. 120–127.
61. Ferrari, T.; Cattaneo, D.; Gini, G.; Golbamaki Bakhtyari, N.; Manganaro, A.; Benfenati, E. Automatic Knowledge Extraction from Chemical Structures: The Case of Mutagenicity Prediction. *SAR QSAR Environ. Res.* **2013**, *24*, 365–383. [CrossRef] [PubMed]
62. Morgan, H.L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113. [CrossRef]
63. Landrum, G. RDKit: Open-Source Cheminformatics. 2006. Available online: https://www.Rdkit.Org/ (accessed on 16 September 2025).
64. Dietterich, T.G. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2000; Volume 1857, pp. 1–15, ISBN 978-3-540-67704-8.

65. Todeschini, R.; Ballabio, D.; Grisoni, F. Beware of Unreliable Q^2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *J. Chem. Inf. Model.* **2016**, *56*, 1905–1913. [CrossRef] [PubMed]
66. Komissarov, L.; Manevski, N.; Groebke Zbinden, K.; Schindler, T.; Zitnik, M.; Sach-Peltason, L. Actionable Predictions of Human Pharmacokinetics at the Drug Design Stage. *Mol. Pharm.* **2024**, *21*, 4356–4371. [CrossRef] [PubMed]
67. Netzeva, T.I.; Worth, A.P.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships: The Report and Recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 155–173. [CrossRef]
68. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810. [CrossRef]
69. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against *Tetrahymena pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746. [CrossRef]
70. Marchiandi, J.; Alghamdi, W.; Dagnino, S.; Green, M.P.; Clarke, B.O. Exposure to Endocrine Disrupting Chemicals from Beverage Packaging Materials and Risk Assessment for Consumers. *J. Hazard. Mater.* **2024**, *465*, 133314. [CrossRef]
71. Chakraborty, P.; Bharat, G.K.; Gaonkar, O.; Mukhopadhyay, M.; Chandra, S.; Steindal, E.H.; Nizzetto, L. Endocrine-Disrupting Chemicals Used as Common Plastic Additives: Levels, Profiles, and Human Dietary Exposure from the Indian Food Basket. *Sci. Total Environ.* **2022**, *810*, 152200. [CrossRef]
72. Schaider, L.A.; Balan, S.A.; Blum, A.; Andrews, D.Q.; Strynar, M.J.; Dickinson, M.E.; Lunderberg, D.M.; Lang, J.R.; Peaslee, G.F. Fluorinated Compounds in U.S. Fast Food Packaging. *Environ. Sci. Technol. Lett.* **2017**, *4*, 105–111. [CrossRef]
73. Undas, A.K.; Groenen, M.; Peters, R.J.B.; Van Leeuwen, S.P.J. Safety of Recycled Plastics and Textiles: Review on the Detection, Identification and Safety Assessment of Contaminants. *Chemosphere* **2023**, *312*, 137175. [CrossRef]
74. Calafat, A.M.; Wong, L.-Y.; Ye, X.; Reidy, J.A.; Needham, L.L. Concentrations of the Sunscreen Agent Benzophenone-3 in Residents of the United States: National Health and Nutrition Examination Survey 2003–2004. *Environ. Health Perspect.* **2008**, *116*, 893–897. [CrossRef] [PubMed]
75. Han, C.; Lim, Y.-H.; Hong, Y.-C. Ten-Year Trends in Urinary Concentrations of Triclosan and Benzophenone-3 in the General U.S. Population from 2003 to 2012. *Environ. Pollut.* **2016**, *208*, 803–810. [CrossRef] [PubMed]
76. Arya, S.; Dwivedi, A.K.; Alvarado, L.; Kupesic-Plavsic, S. Exposure of U.S. Population to Endocrine Disruptive Chemicals (Parabens, Benzophenone-3, Bisphenol-A and Triclosan) and Their Associations with Female Infertility. *Environ. Pollut.* **2020**, *265*, 114763. [CrossRef] [PubMed]
77. Mansouri, K.; Kleinstreuer, N.; Abdelaziz, A.M.; Alberga, D.; Alves, V.M.; Andersson, P.L.; Andrade, C.H.; Bai, F.; Balabin, I.; Ballabio, D.; et al. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environ. Health Perspect.* **2020**, *128*, 27002. [CrossRef]
78. Benfenati, E.; Manganaro, A.; Gini, G. VEGA-QSAR: AI inside a Platform for Predictive Toxicology. In *CEUR Workshop Proceedings Vol-1107, Proceedings of the Popularize Artificial Intelligence 2013, Turin, Italy, 5 December 2013*; CEUR-WS: Aachen, Germany, 2013.
79. Manganelli, S.; Roncaglioni, A.; Mansouri, K.; Judson, R.S.; Benfenati, E.; Manganaro, A.; Ruiz, P. Development, Validation and Integration of in Silico Models to Identify Androgen Active Chemicals. *Chemosphere* **2019**, *220*, 204–215. [CrossRef]
80. Triebe, J.; Worth, A.; Janusch Roi, A.; Coe, A. *JRC QSAR Model Database: EURL ECVAM DataBase Service on Alternative Methods to Animal Experimentation: To Promote the Development and Uptake of Alternative and Advanced Methods in Toxicology and Biomedical Sciences: User Support & Tutorial, EUR 28713 EN*; Publications Office of the European Union: Luxembourg, 2017; JRC107491, ISBN 978-92-79-71406-1. [CrossRef]
81. Watanabe, J.; Kozaki, A. Relationship between Partition Coefficients and Apparent Volumes of Distribution for Basic Drugs. II. *Chem. Pharm. Bull.* **1978**, *26*, 3463–3470. [CrossRef]
82. Hähnke, V.D.; Kim, S.; Bolton, E.E. PubChem Chemical Structure Standardization. *J. Cheminform.* **2018**, *10*, 36. [CrossRef]