JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

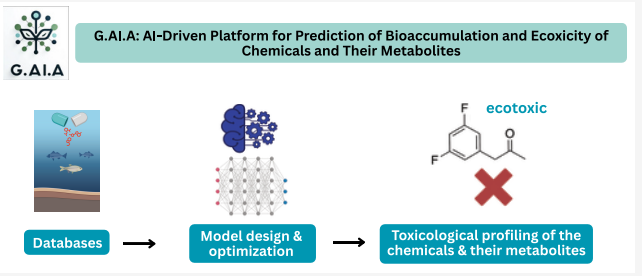
pubs.acs.org/jcim Article

G.AI.A: An Integrated Machine-Learning Platform for Predicting Bioaccumulation and Ecotoxicity of Pharmaceuticals

Evangelos Tsoukas,[‖] Michail Papadourakis,[‖] Eleni Chontzopoulou, Spyridon Vythoulkas, Christos Didachos, Dionisis Cavouras, Panagiotis Zoumpoulakis,* and Minos-Timotheos Matsoukas*

Cite This: https://doi.org/10.1021/acs.jcim.5c02286 Read Online

ABSTRACT: Pharmaceutical pollution in aquatic environments poses a significant ecological threat due to the accumulation of bioactive compounds from human and veterinary sources. In support of the EU Green Deal's Chemicals Strategy for Sustainability, this study presents a computational framework for predicting two key environmental risk indicators in fish: bioconcentration and ecotoxicity. Bioconcentration, quantified by the bioconcentration factor (BCF), reflects a chemical's tendency to accumulate in organisms, while ecotoxicity is assessed via the median lethal concentration (LC_{50}) over defined exposure periods. We developed two high-performing machine learning (ML) models, achieving ROC AUC scores of 94.60% for bioconcentration and 96.06% for ecotoxicity, validated across both internal and external data sets. To expand the scope of risk evaluation, we incorporated metabolite prediction using the SyGMa tool, selected after benchmarking multiple alternatives. This enables the assessment of both parent compounds and their potentially toxic metabolites. Model interpretability was enhanced through molecular fingerprint analysis, which identified structural features associated with toxicity and accumulation, informing the early stages of drug design. To support practical implementation, we introduced G.AI.A (https://gaiatox.eu/), an intuitive web platform that allows users to input Simplified Molecular Input Line Entry System (SMILES) strings for rapid prediction of environmental risk end points. The application domain of G.AI.A lies in predictive toxicology, enabling researchers and regulatory bodies to assess the toxicological profiles of small organic compounds, excluding those containing heavy metals, by analyzing their chemical structures. The platform supports batch processing and offers interactive visualizations, facilitating compound screening and early stage environmental risk assessment. By integrating predictive modeling with interpretability and usability, our framework advances green-by-design pharmaceutical development and contributes to sustainable chemical management.



INTRODUCTION

The Chemicals Strategy for Sustainability (CSS) is a key component of the EU Green Deal, aiming to reduce hazardous chemical use in line with Europe's goal of climate neutrality by 2050.[1−3] As part of this effort, regulations such as Classification, Labeling and Packaging of Hazardous Substances (CLP),[4] the EU Regulation on the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH),[5] and the Restriction of Hazardous Substances Directive[6] are being revised to limit the use of dangerous substances unless no safer alternatives exist.[7] Among the targeted substances, pharmaceuticals pose a growing environmental concern, as large quantities accumulate in aquatic ecosystems due to human and veterinary use. Human pharmaceuticals enter waterways through wastewater treatment plants, while veterinary drugs reach the environment via runoff, manure, and aquaculture.[8,9] These contaminants threaten ecosystems, highlighting the urgent need to mitigate their environmental impact.[10]

In evaluating the environmental risk of pharmaceuticals, it is crucial to measure both their bioconcentration and ecotoxicity to determine their overall impact on ecosystems. Bioconcentration, a measure of a chemical's ability to accumulate in living organisms, is quantified using the bioconcentration factor (BCF) which is a vital metric for assessing a substance's potential to accumulate in aquatic organisms, calculated as the ratio of the steady-state concentration within the organism to that in the surrounding water. Substances with BCF values between 2000 L/kg (3.30 log units) and 4999 L/kg (3.699 log units) are considered "bioaccumulative". Ecotoxicity, on the other hand, is often evaluated using the Lethal Concentration (LC_{50}), which indicates the concentration of a substance in a

Received: September 25, 2025
Revised: December 30, 2025
Accepted: December 31, 2025

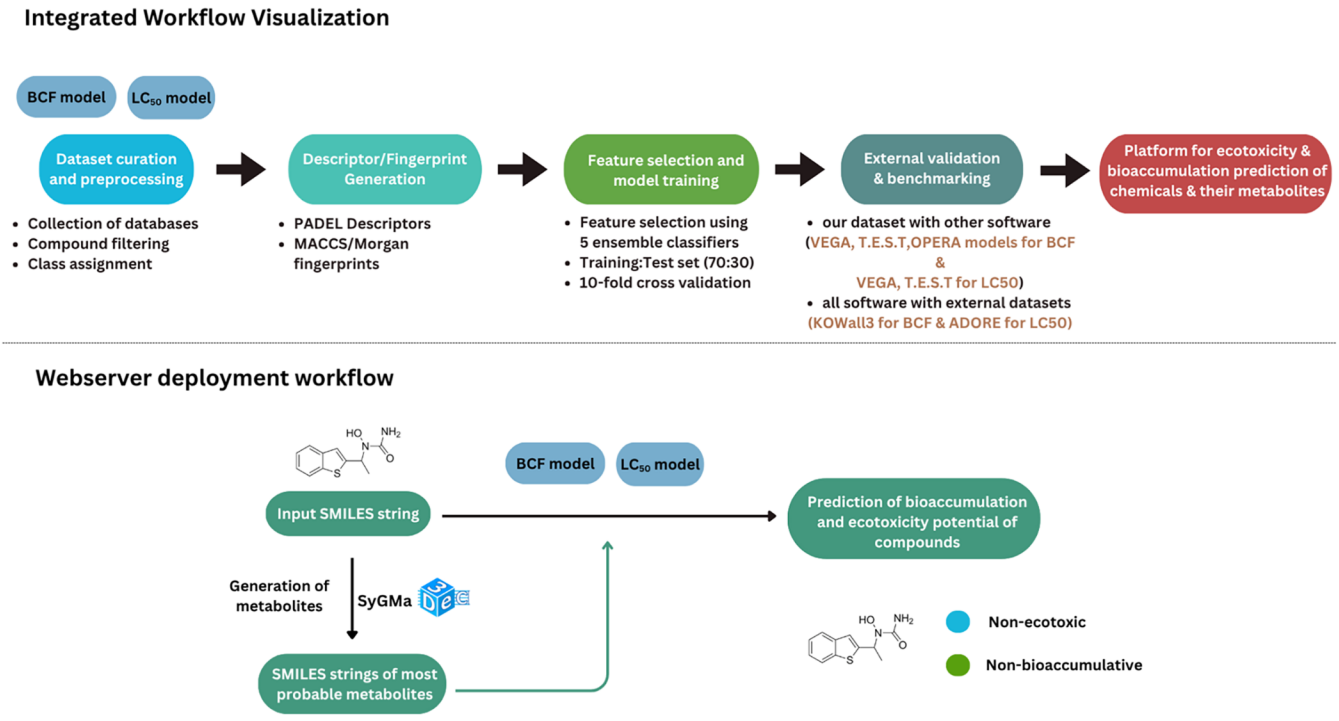


Figure 1. Overview of the workflow employed in this paper.

given medium (e.g., water) that is lethal to 50% of the test organisms over a specified exposure period (e.g., 96 h). According to CLP regulations, substances with LC_{50} values below 1 mg/L in 96-h tests are classified as "ecotoxic".

Various computational attempts have been made to develop predictive models for the BCF of chemical substances in aquatic organisms, particularly fish species. Banjare et al. created a Quantitave Structure Activity Relationship (QSAR) model for 55 pesticides,[11] while Lunghini et al. used ISIDA descriptors and machine learning (ML) for industrial chemicals.[12] Sushko et al. utilized the OCHEM[13] platform to calculate 2D descriptors and applied regression-based quantitative structure−property relationship modeling further enhancing prediction by combining the models through the "intelligent consensus" algorithm to identify chemical features associated with acute fish bioconcentration.[14,15] Kobayashi & Yoshida developed an ML-based QSAR model for 522 compounds,[16] and Zhao et al. proposed a deep learning model with R^2 values of 0.70 (training) and 0.68 (test).[17] Xu et al. developed a Quantitative Structure *In vitro-In vivo* Relationship model, optimized with ML algorithms, to predict BCF values for multiple chemical substances and species.[18] Finally, Pore et al. utilized the quantitative read-across approach,[19] while Chen et al. developed a multitask deep learning model to predict BCF for various organic compounds.[20] In addition, many *in silico* methods have been developed to predict the ecotoxicity of chemical compounds in aquatic organisms, offering the potential to supplement or even replace animal testing by leveraging predictive toxicology based on historical data.[21] Traditional QSAR models like ecological structure−activity relationship (ECOSAR[22]) and Toxicity Estimation Software Tool (T.E.S.T[23]) have long been used to estimate toxicity across species such as algae, daphnia, and fish using linear regression methods. Recent advancements in QSAR models have focused on improving the accuracy of predicting organic chemical toxicity in ecologically important fish species, utilizing both traditional approaches[24] and ML techniques.[25−27] However, the focus has increasingly shifted toward ML approaches in recent developments. Li et al. utilized ML models to predict the acute toxicity of pesticides to fish, achieving a balanced accuracy of 0.83 on a small data set.[28] Tuulaikhuu et al. used EcoTox data with taxonomy features and random forests, reaching R^2 = 0.85 for acute fish toxicity.[29] Wu et al. combined ML and Read Across Structure Activity Relationships (RASAR) models, predicting fish toxicity with over 93% accuracy.[30] Similar approaches have used both conventional ML algorithms and graph convolutional networks, achieving classification accuracies exceeding 83% in predicting fish toxicity.[31] Viljanen et al. tested several ML methods on ECOTOX data (2,431 chemicals, 1,506 species) to predict LC_{50} values.[32] Finally, Gasser et al. evaluated ML models in the "t-F2F" challenge[33] using six molecular representations, with the best model reaching RMSE equal to 0.90.[34] While QSAR models have significantly improved the prediction of chemical bioconcentration and ecotoxicity, they often fall short in terms of user-friendliness, such as limitations on the number of compounds that can be input[35] or outputs being presented in a disorganized way, making it hard to link inputs to results.[22] Many models also lack transparency, offering no access to underlying code or web services,[36−40] and suffer from poor interpretability due to unclear feature selection.[41] Additionally, their applicability to diverse chemical data sets remains limited, as external validations are frequently performed on small or narrow data sets.[16,37,40,42] "Moreover, despite the strong predictive performance of ML and deep learning approaches, their "black-box" nature hampers transparency and interpretability, as it is difficult to trace the contribution of individual chemical features to model predictions. This limitation reduces their suitability for regulatory use, where understanding the drivers of predictions is essential for ensuring environmental safety.

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

In the evolving landscape of drug development and environmental safety assessment, accurate prediction of bioconcentration and toxicity is crucial not only for the parent drug but also for its metabolites, as these derivatives can play a pivotal role in shaping the overall safety profile of pharmaceutical compounds. This human metabolic process involves two main routes: an initial step featuring oxidation, reduction, hydrolysis, and alkylation reactions (Phase I metabolization), and a subsequent pathway wherein conjugates, primarily of glucuronide or sulfate types, are predominantly formed (Phase II metabolization). Ultimately, these compounds are excreted in urine or feces, predominantly as conjugates, presenting as more polar and hydrophilic derivatives. This excretion commonly occurs as either a single major metabolite or, more frequently, as a combination of multiple metabolites.[43] The European Medicines Agency (EMA) mandates reporting total concentrations of pharmaceuticals, including their metabolites, to provide a comprehensive view of their environmental footprint.[44,45] Additionally, assessing the bioaccumulation and ecotoxic potential of these metabolites is crucial for a comprehensive understanding of the drug's safety profile.

In this study, we present the development of interpretable ML models for predicting the bioconcentration and ecotoxicity of organic chemicals in fish. We examined the structural features and key physicochemical descriptors associated with compounds exhibiting high toxicity or bioaccumulation potential. Beyond building state-of-the-art *in silico* tools for forecasting the ecotoxic effects of pharmaceuticals (i.e., parent compounds), we also assessed publicly available software to identify potential metabolites and transformation products (e.g., those with high probability to occur in the aquatic environment). The final objective of this step was to evaluate the bioconcentration and ecotoxicity profiles of derivatives predicted using these computational tools. Hence, the most efficient tools were ultimately integrated into an automated pipeline that facilitates a thorough evaluation of pharmaceutical ecotoxicity by assessing both parent compounds and their metabolites, ensuring compliance with green chemistry principles. This approach provides computational workflows and filtering mechanisms to support green drug design at early development stages, enabling the identification of environmentally safer molecules. However, it may not fully capture toxicity in ecological contexts or for species outside the training domain. An overview of the employed workflow is illustrated in Figure 1.

METHODS

Data Generation and Description

In this study, both the bioaccumulation and toxicity data sets were compiled from experiments conducted using standardized test methods or internationally recognized guidelines (e.g., OECD), ensuring that the underlying data are reliable and comparable. For bioaccumulation, we only included BCF measurements derived from fish species recommended by OECD guidelines.[46] Due to the limited availability of experimental BCF values, BCF data for 1,243 unique organic compounds tested on fish species were compiled from various public databases and literature sources, including Lunghini et al.,[12] Miller et al.,[47] the US EPA's ECOTOX,[18] and Yuan et al.[48] The data set was organized into four groups based on bioconcentration in different fish species: *Cyprinus*, *Oncorhynchus*, nine additional species, and a combined group encompassing 54 different fish species. Due to variability in data distribution, the median of the median logBCF values was used to classify compounds according to the REACH threshold of logBCF > 3.30 log units (2000 L/kg).[49] Simplified Molecular Input Line Entry System (SMILES) strings were generated using Open Babel v2.4.1,[50] and additional curation and standardization procedures, including filtering, deduplication, and structure validation, are detailed in the Supporting Information (SI). For external validation, we used the KOWall3 data set from Xiao et al. which contains bioaccumulation parameters for 1,724 compounds from various databases and literature sources.[51] Similarly, the LC_{50} data set was constructed following standardized acute toxicity protocols, primarily OECD Test Guideline 203,[52] which specifies a 96-h exposure period with mortality recorded every 24 h. Because of the scarcity of high-quality experimental LC_{50} values, LC_{50} data for 2,940 compounds were sourced from the EnviroTox[53] and OPERA[38] databases, with two external ecotoxicity validation sets (t_F2F and s-F2F-2) derived from the curated ADORE benchmark. Each compound was classified based on the experimentally observed minimum 96-h LC_{50} value, using the CLP ruleset threshold (LC_{50} = 1 mg/L). We reviewed the data accessibility of VEGA, OPERA, and T.E.S.T., incorporating available external data sets (VEGA BCF CAESAR,[54] OPERA BCF,[55] and VEGA Fathead Minnow LC_{50}[56]). SI contains a full description of this evaluation. Finally, t-SNE plots comparing our data sets with external data sets (VEGA BCF CAESAR and OPERA for BCF models and VEGA Fathead Minnow for LC_{50} model), as well as with the KOWall3 and ADORE data sets, are presented in the Supporting Information (Figures S1). It should be noted that all publicly available data sets used here have been previously applied in other ML studies, as noted in the Introduction. Also, despite our efforts to ensure comprehensive data collection, information on certain key environmental factors (e.g., pH, temperature) was not consistently available across studies due to differences in experimental design and reporting limitations. While this limitation may somewhat affect our understanding of bioconcentration or ecotoxicity parameters, we believe that the compiled data sets still provides valuable insights for exploring bioaccumulation and ecotoxicity.

Feature Selection and Models Construction

Three types of molecular encodings were explored for training and evaluating the ML models: 1D, 2D, and 3D molecular descriptors (calculated using PaDEL,[57] version 0.1.16), along with Molecular ACCess System keys (MACCS[58]), and Morgan[59] fingerprints generated via RDKit.[60] Morgan fingerprints were produced using radii of 2 to 4 and bit lengths of 1024 and 2048. The data sets were preprocessed by eliminating descriptors with zero variance, low variance (σ < 0.1), high correlation (r > 0.80), or missing values (NaNs). Numerical descriptors were normalized using Scikit-learn's StandardScaler, while binary fingerprints were excluded from scaling. Five ensemble classifiers, RandomForest,[61] ExtraTrees,[62] XGBoost,[63] AdaBoost,[64] and GradientBoosting,[65] were used to classify compounds as bioaccumulative/ecotoxic (class 1) or nonbioaccumulative/nonecotoxic (class 0). Feature selection was conducted by training each model for 200 epochs on randomly scaled 70% subsets of the data, identifying the 20 most important features per descriptor or fingerprint type based on the feature rankings provided by the Random Forest, Extra Trees, XGBoost, AdaBoost, and Gradient Boosting classifiers. Final model optimization was carried out by incrementally increasing the number of features from 5 to 20, selecting configurations with the highest area under the receiver-operating characteristic curve (ROC AUC) scores from both internal cross-validation and external validation sets. Full details on feature processing, model tuning, and validation are available in the SI.

Evaluation of Model Performance

To prepare the data set for model training, we performed a 70:30 split into training and test sets using Scikit-learn's train_test_split. The 70% subset was used exclusively for model training and internal validation, while the remaining 30% was held out as an independent external test set to evaluate predictive performance on unseen compounds, particularly for BCF and LC_{50} responses. Since both data sets were imbalanced, with a higher proportion of nonbioaccumulative and nonecotoxic chemicals, oversampling techniques, including

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Overview of the Freely Available Predictive Tools

tool/platform	VEGA		OPERA	T.E.S.T.	
model type/ algorithm	multiple algorithms including BCF models (e.g., CAESAR), LC_{50} models (e.g., IRFMN-Combase), and others		multiple models, primarily BCF models; LC_{50} models not included	multiple models covering both BCF and LC_{50} end points, plus various clustering methods	
interpretability	provides applicability domain index (ADI), similar compounds, structural alerts, and chemical reasoning		transparent feature selection, interpretable descriptors, and similar compounds displayed	shows training set analogs, cluster information, fragment analysis, and prediction intervals	
transparency (code access)	open source − models available, integrated in QSAR Toolbox, QMRF documents available; data sets not downloadable		open source − GitHub repository (kmansouri/OPERA), full code and data sets available	open source − EPA-provided, full methodology documented in the user guide, data not downloadable	
web service availability	yes − VEGAHUB web interface, standalone desktop application (Java), batch processing supported		yes − via CompTox Dashboard API, command-line interface, and integration in other platforms	yes − WebTEST 2.0 platform with REST API and standalone GUI application	
	BCF model (CAESAR)	LC_{50} toxicity model (IRFMN-Combase)	BCF	BCF (Consensus clustering)	96-h fathead minnow LC_{50} (Consensus clustering)
performance metrics	R^2 = 0.78, RMSE = 0.62	R^2 = 0.73, RMSE = 0.72	R^2 = 0.83, RMSE = 0.64	R^2 = 0.75, RMSE = 0.67, MAE = 0.52	R^2 = 0.73, RMSE = 0.77, MAE = 0.55

SMOTE[66] and SMOTENC for binary fingerprints, were employed to balance class distributions and augment the training data. In smaller fish-specific data sets, we additionally applied resampling with added Gaussian noise[67] to enhance model robustness.[68] Models were trained using 7-fold cross-validation repeated 50 times via RepeatedStratifiedKFold to minimize overfitting and ensure generalizability. Final model evaluation was conducted on a fully independent test set using metrics such as accuracy, sensitivity, specificity.[69] Additionally, ROC curves were generated, and the ROC AUC values were computed to further evaluate model performance. To interpret model predictions, the SHAP (SHapley Additive exPlanations)[70] method was used to quantify feature importance, with a threshold of 0.1 for significant contributors. SHAP insights were combined with 2D PCA[71] plots of PaDEL descriptors and statistical analyses (boxplots distributions, Mann−Whitney U test[72]) to identify key features associated with bioaccumulation and ecotoxicity. For the MACCS- and Morgan-based ML models, we calculated the percentage of compounds per class that exhibited a value of 1 for each training feature. SHAP analysis was then applied to determine the influence of class 1 prevalent structural bits on overall model decision-making. The identified substructures can be interpreted as chemical groups associated with bioconcentration and ecotoxicity, providing insights for environmental risk assessment in drug design. Full details of the preprocessing, cross-validation, interpretability workflow, and statistical analyses are provided in the SI.

Our curated logBCF and external validation logBCF data sets, along with our LC_{50} and external validation LC_{50} data sets, served as benchmarks for assessing our best predicted BCF model against the freely available tools: Virtual models for property Evaluation of chemicals within a Global Architecture (VEGA), Open (Quantitative) Structure−Activity/Property Relationship App (OPERA), and T.E.S.T. Similarly, our best − performing LC_{50} model was evaluated against VEGA and T.E.S.T. Comprehensive details of the software tools considered in this study are included in the SI, with an overview summarized in Table 1.

Prediction of Drugs' Metabolites

To identify suitable tools for predicting human drug metabolites, we reviewed publicly available software capable of generating metabolite structures from parent compounds, excluding those limited to predicting only sites of metabolism (SoMs). Three tools, SyGMa,[73] Metapredictor,[74] and Metatrans,[75] were selected based on their use of rule-based or ML methods and their ability to output complete metabolite structures. These tools were evaluated using a validation set of 221 drugs with experimentally verified human metabolites sourced from the ChEMBL database.[76] To assess predictive accuracy, we calculated the ratio of experimentally verified to total predicted metabolites. Tool-specific settings, such as beam size for Metatrans and Metapredictor, were optimized to balance prediction breadth and relevance. While Biotransformer 3.0[77] was reviewed, it was excluded from performance evaluation due to excessive metabolite generation. However, despite its exclusion, we still included enviPath,[78] an ML-based algorithm integrated into Biotransformer, in our assessment. EnviPath predicts microbial biotransformations of organic environmental contaminants and metabolic products from both abiotic and biotic aquatic processes. Full methodological details, software configurations, and evaluation metrics are provided in the SI.

RESULTS AND DISCUSSION

Bioconcentration Factor Data Quality and Variability

A total of 1243 experimental logBCF data points for unique chemicals associated with fish species were collected from four databases. Figure S2 presents the distribution of logBCF values (ranging from −2.05 to 5.97 log units) alongside three key chemical properties, molecular weight, water solubility, and logP, highlighting the broad applicability domain of the data set used for model training and validation. The BCF data were divided into four data sets based on the bioconcentration of chemical compounds in different fish species: *Cyprinus*, *Oncorhynchus*, nine other fish species, and all fish species combined. For the *Cyprinus* data set, based on the REACH regulatory logBCF > 3.30 log units (2000 L/kg) cutoff, 95 compounds were classified as bioaccumulative (class 1) and 610 as nonbioaccumulative (class 0). For the *Oncorhynchus* data set, 89 compounds were identified as bioaccumulative (class 1) and 140 as nonbioaccumulative (class 0). Regarding the nine other fish species available, 279 compounds were categorized as bioaccumulative (class 1) and 796 as non-bioaccumulative (class 0). Finally, for the combined data set, 214 compounds were designated as bioaccumulative (class 1) and 1054 as nonbioaccumulative (class 0).

Model Performance on Bioaccumulation Classification

To establish a solid foundation for our methodology development, we initially concentrated our efforts on a single fish species using the PaDEL descriptors. The *Cyprinus* species was chosen due to having the highest number of compounds with reported experimental logBCF data (705 compounds), followed by the *Oncorhynchus* data set, which contained the second-largest number of reported experimental logBCF data

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

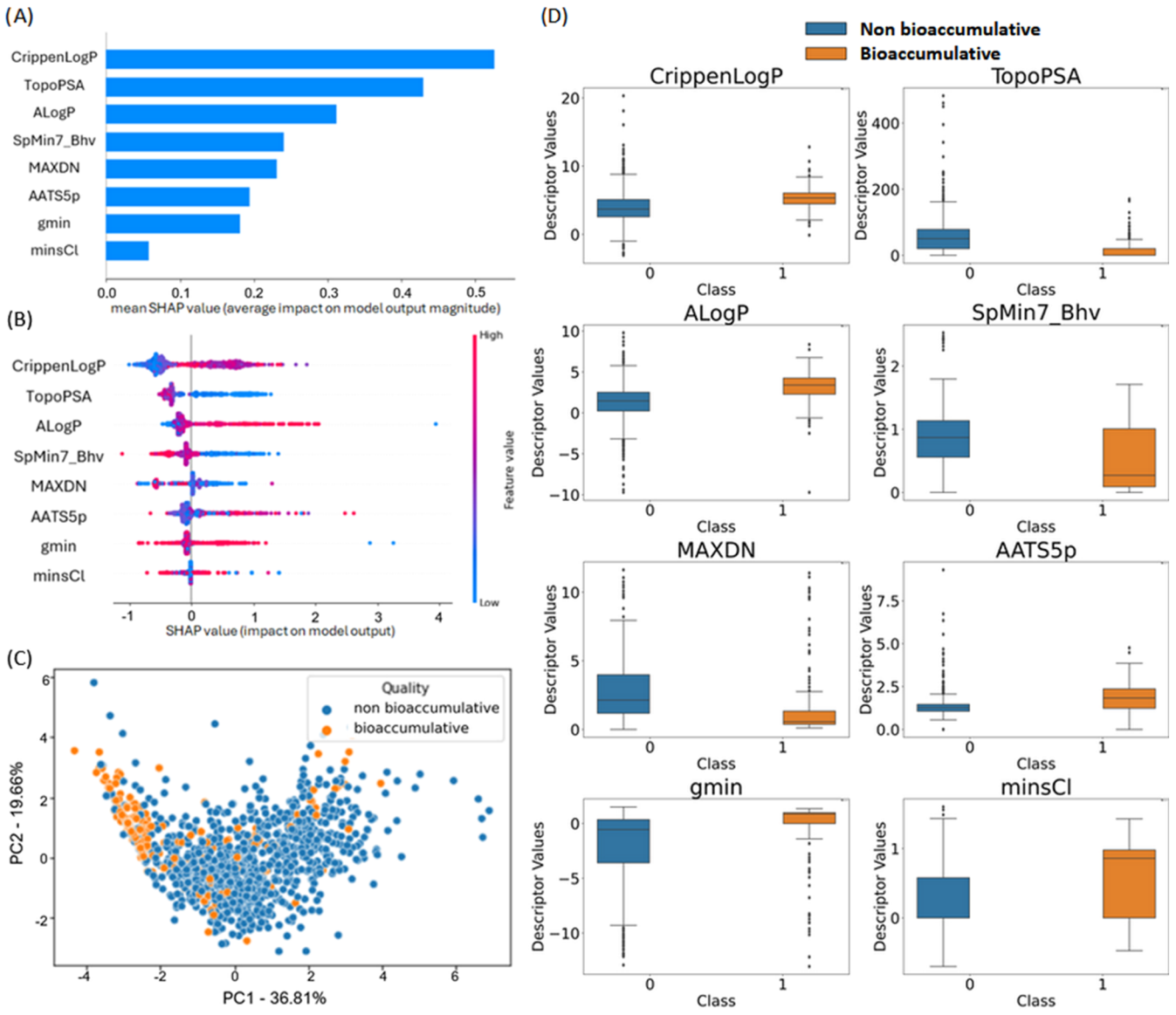


Figure 2. (A) Histogram of mean absolute SHAP values for the 8 features in differentiating bioaccumulative compounds. (B) Distribution of local SHAP values for the most influential descriptors in differentiating bioaccumulative compounds. (C) 2D PCA plot depicting the training features of the bioconcentration ML model. (D) Boxplots representing the distribution values of the 8 features in distinguishing bioaccumulative compounds across the two bioconcentration classes.

(229 compounds). The generated models (Tables S1 and S2) were quite efficient in distinguishing bioaccumulative and nonbioaccumulative compounds in both *Cyprinus* (best reported accuracy = 93.55% ± 0.80%) and *Oncorhynchus* data sets (best reported accuracy = 96.46 ± 0.88%). Based on the high predicted performance of the 5 top-scoring models in terms of accuracy, sensitivity and specificity we decided to generalize our models by including 9 other fish species (1,075 number of compounds). The top-performing models demonstrated outstanding results, with accuracy, specificity, and sensitivity all surpassing 85% (Table S3). As a result, we broadened the applicability of our approach to include the entire data set, which consists of 54 different fish species and a total of 1268 compounds. Among the models tested (Table S4), Gradient Boosting emerged as the top performer in terms of ROC AUC (86.95 ± 2.69%, Figure S3). It also outperformed Random Forest, XGBoost, and ExtraTrees in sensitivity (74.50 ± 5.97%), while achieving comparable sensitivity to AdaBoost. This suggests that the algorithm can adequately classify bioaccumulative and nonbioaccumulative compounds. Due to its overall superior performance and generalization across all 54 different fish species, this model was selected to be integrated into the G.AI.A platform.

In this study, a SHAP analysis was performed on the Gradient Boosting model to assess the importance of its features, as detailed in the Methods section (Table S5 and Figure 2). Positive SHAP values (greater than 0.1) indicate stronger influence in differentiating between bioaccumulative and nonbioaccumulative chemical compounds. In addition, the Mann−Whitney *U* test was applied as a postprocessing step to rank the 8 features (definition of their types is provided in Table S5). Calculated *P* values from this test showed that these features have significantly different distributions between bioaccumulative and nonbioaccumulative compounds, providing a chemical perspective of the binary classification process. "CrippenLogP", "TopoPSA", and "ALogP" were significantly

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

more important than the next features. High values of "CrippenLogP" and "ALogP" (Figure 2B) which correspond to increased lipophilicity, result in large mean SHAP values (Figure 2A), suggesting that compounds with greater lipophilicity tend to be more bioaccumulative. In contrast, an inverse trend is observed for "TopoPSA", indicating that bioaccumulative compounds in fish generally have smaller polar surface areas. These findings align with the principle that compounds with higher lipophilicity tend to accumulate more in fatty tissues, leading to greater bioaccumulation in fish species. The same applies to "AATS 5p", a measure of correlation between the polarizability of atoms separated by five bonds in the molecular structure. Polarizability, which is also linked to the hydrophobicity of compounds, has been confirmed as a highly influential factor in the mechanism of bioconcentration.[16,47,79−82] However, it is important to note that chemicals with low polarizability may still accumulate through alternative mechanisms, such as protein binding.[83] "gmin" and "MAXDN", which are linked to the molecule's electronic properties may be related to polarity-related accumulation across cellular membranes.[84] Based on the observed trends, bioaccumulative compounds in fish generally exhibit lower nucleophilicity and feature atoms with higher minimum electron density. The patterns described above are clearly reflected in the boxplots of these features, highlighting the features used by the model across the two classes in the multispecies fish data set (Figure 2D). For "SpMin7_Bhv,", which is directly associated with molecular size, and "minsCL", reflecting the minimum electron density around any chlorine atom in the molecule, the boxplot distributions do not allow for a clear or straightforward interpretation. Finally, a 2D PCA plot was generated using the features of the Gradient Boosting model presenting that the previously mentioned features were instrumental in distinguishing the class 1 cluster (Figure 2C).

The models based on PaDEL descriptors provided valuable insights into the physicochemical properties of bioaccumulative compounds. However, it was essential to extend our research to include featurization using molecular fingerprints. Therefore, we analyzed the multispecies fish data set using Morgan (with varying radii and bit lengths) and MACCS fingerprints as training features to identify the structural characteristics of bioaccumulative compounds and to explore the development of a more efficient ML model. Table S6 presents the results of the most efficient ML models after 10 epochs of external validation on random test sets. The Morgan-based ML models generated from the corresponding bioconcentration data set showed lower accuracy compared to the most efficient descriptor-based model (ROC AUC of 86.95% ± 2.69%). Notably, among the Morgan-based ML models, the highest predictive performance, with a ROC AUC of 79.20 ± 2.99%, was achieved using the ExtraTrees algorithm with Morgan fingerprints of radius 2 and 2048 bits as training features. In contrast, the MACCS-based ML model showed improved predictive performance, similar to the descriptor-based model, with a ROC AUC of 84.37 ± 2.59%. The SHAP analysis of the top-performing Morgan- and MACCS-based models, which aided in identifying structural alerts relevant for regulatory purposes, is presented in the SI and visualized in Figures S4−S7. In the best MACCS-based ML model (which demonstrated similar performance to the best descriptor-based model), bit 87, representing halogen-containing fragments, had the second-highest mean SHAP value (Figure S6A) and was present in 70% of the bioaccumulative class, compared to only 30% in the nonbioaccumulative class. Similarly, bit 103, which also indicates the presence of chlorine atoms, had a lower impact on distinguishing bioaccumulative from nonbioaccumulative compounds but was still found in 67% of class 1. These findings align with previous research indicating that the presence of chlorine atoms enhances hydrophobicity[85] and contributes to the environmental persistence of chemical compounds.[14] Finally, bit 62, which represents polycyclic aromatic fragments, had half the mean SHAP value of bit 87. Although it was present in only 30% of bioaccumulative compounds, it appeared in just 10% of the nonbioaccumulative class, indicating that these fragments are predominantly associated with bioaccumulative compounds. Compounds containing these fragments are often resistant to biodegradation,[86] resulting in prolonged environmental persistence and accumulation within the food chain. Structural features such as halogen-containing fragments and polycyclic aromatic systems can increase a compound's lipophilicity, facilitating its absorption and retention in the lipid-rich tissues of aquatic organisms. Additionally, these substructures may exhibit specific binding affinities to proteins and enzymes, further prolonging their retention time and enhancing bioaccumulation potential. For example, the presence of multiple chlorine atoms can strengthen interactions with biological macromolecules, potentially inhibiting metabolic pathways responsible for degradation and elimination.[87] These chemical features have also been recognized as key factors in enhancing the bioaccumulation of compounds in other in silico models.[20,88] Recognizing these structural alerts is crucial for regulatory agencies, as they provide a simple yet effective visualization tool for risk assessment and regulatory decision-making.

Comparison with Previous Studies for Bioaccumulation Classification

To further evaluate our top-performing ML models, a thorough literature review was conducted (described in Introduction) to identify publicly available tools for predicting the BCF of chemical compounds in aquatic organisms, enabling a performance comparison between our G.AI.A models and these tools using our curated logBCF data set. A comprehensive benchmarking study of our proposed bioconcentration model was performed, comparing its performance with other publicly available tools, including VEGA,[89] OPERA,[55] and T.E.S.T.[23] All models were evaluated using the same data set of molecular compounds compiled for the development of G.AI.A models (Table S7). Although all models exhibited high accuracy and specificity, their performance on sensitivity, the key metric for distinguishing the minority class in our data set, showed considerable variation. VEGA and OPERA showed the lowest sensitivity (20.00% and 17.43%, respectively). T.E.S.T achieved a sensitivity of 70.58%, comparable to our model's performance (74.50% ± 5.97), but its ROC AUC of 75% was markedly lower, highlighting the superior stability and robustness of our model (ROC AUC = 86.95% ± 2.69). In addition, we conducted further benchmarking of our model against the three software tools using the two publicly available external data sets from VEGA (VEGA BCF CAESAR) and OPERA (OPERA BCF), with the results provided in Tables S8 and S9. For the VEGA BCF CAESAR data set, our model achieved higher values across most performance metrics, with specificity being the only metric where it was slightly lower (90.91% versus 97.92, 95.12,

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

and 98.67% for VEGA, OPERA, and T.E.S.T., respectively). Notably, our model demonstrated superior sensitivity (88.51% versus 62.37, 74.19, and 61.29% for VEGA, OPERA, and T.E.S.T.). For the OPERA BCF data set, OPERA achieved perfect scores across all metrics, which is expected when a model is evaluated on its own data set but also suggests a narrow applicability domain. T.E.S.T. also achieved 100% sensitivity but performed poorly in overall accuracy (20.28%) and specificity (2.56%). Our model produced balanced and consistent performance across metrics, with a moderate decrease in sensitivity (64.00%), yet still outperforming VEGA (50.00%), demonstrating the generalizability of our approach. Finally, the other studies and tools place less emphasis on model interpretability and often provide limited chemical insight into their predictions. In contrast, we sought to offer an understanding of the physicochemical properties and chemical groups linked to bioconcentration.

Furthermore, we conducted a comprehensive benchmarking study of our proposed bioconcentration model, assessing its performance against the same tools using an external data set of molecular compounds from Xiao et al. (Table 2). Figure S8

Table 2. Validation Metrics Obtained from Testing Publicly Available Online Software Using the Xiao et al. logBCF Dataset

online software/model	accuracy (%)	specificity (%)	sensitivity (%)	ROC AUC (%)
VEGA	88.72	98.25	50.76	87.98
OPERA	90.57	93.67	78.13	93.66
T.E.S.T	88.37	97.58	56.40	92.83
(our) G.AI.A MODEL	88.60	89.73	83.87	93.80

presents a comparison of the chemical space covered by our curated logBCF data set (BCF_data set) and the external logBCF data set from Xiao et al., which includes multiple fish species (external_BCF_data set). This comparison highlights the broad applicability domain of our data set relative to commonly used external validation sets. Notably, 146 compounds in the Xiao data set are already included in our logBCF data set and were part of the model training. In the Xiao data set, 31 compounds were not processed by PaDEL software, and 262 compounds were also excluded due to their low molecular weight (MW < 150 Da). All models demonstrated comparable accuracy on this data set, achieving higher performance than on our logBCF data set, with accuracy ranging from 88.37 to 90.57%. Notably, all publicly available tools outperformed our G.AI.A model (89.78%) in terms of specificity, with VEGA achieving the highest specificity (98.25%), followed by T.E.S.T (97.58%) and OPERA (93.67%). However, our G.AI.A model demonstrated markedly higher sensitivity (84.88%), substantially outperforming the top existing models (VEGA: 50.76%, T.E.S.T.: 56.40%, and OPERA: 78.13%). Moreover, our G.AI.A model achieved the highest ROC AUC score (94.60%), surpassing OPERA (93.66%), T.E.S.T (92.83%), and VEGA (87.98%), demonstrating its superior stability and robustness.

Ecotoxicity Data Quality and Variability

A total of 2940 experimental acute LC_{50} data points for unique chemicals associated with fish species were collected from two databases (Envopera). Figure S9 presents the distribution of $logLC_{50}$ values (ranging from − 6.00 to 4.76 mg/L) alongside three key chemical properties: molecular weight, water solubility, and logP highlighting the broad applicability domain of the data set used for model training and validation. The data set included multiple experiments for the same chemicals with differing LC_{50} values; hence, the minimum LC_{50} value for each chemical was selected for classification based on the CLP ruleset, using an LC_{50} < 1 mg/L threshold. This classification identified 1,991 compounds as nonecotoxic (class 0) and 1,014 as ecotoxic (class 1).

Model Performance on Ecotoxicity Classification

The same five classifiers and feature-selection pipeline used for the BCF analysis were applied to the LC_{50} data set, using the same training/test split and 10-fold cross-validation. As presented in Table S9, the Random Forest, XGBoost, and ExtraTrees algorithms achieved accuracies exceeding 85%, outperforming ADABoost (76.84% ± 1.13) and Gradient Boosting (75.36% ± 0.68). Notably, Random Forest and ExtraTrees achieved specificity and ROC AUC metrics exceeding 90%, while also demonstrating solid performance in identifying ecotoxic compounds, with sensitivities of 74.98% ± 3.73 and 73.38% ± 3.16, respectively. Finally, Figure S10 showcases the predictive performance of our top-performing ExtraTrees model in terms of specificity, evaluated across ten epochs of external validation on random test sets, with ROC AUC serving as the evaluation metric. Due to its overall superior performance and generalization across all fish species, this model was selected to be integrated into the G.AI.A platform.

The importance and contribution of each feature to the performance of the ExtraTrees model were again assessed through SHAP analysis. The full set of 14 features used to train the model is provided in Table S10, with their respective SHAP analysis illustrated in Figure S11. The 8 most impactful features for distinguishing ecotoxic compounds in fish are summarized in Table S11. Among these, "CrippenLogP" and "ALogP", indicators of increased lipophilicity, were pivotal in the performance of the top-ranking bioconcentration ML model and demonstrated high mean absolute SHAP values, underscoring their significant contribution to the ecotoxic ML model's predictive capabilities. These findings align with the ecotoxicological principle that compounds with higher lipophilicity tend to accumulate in fatty tissues, resulting in increased toxicity. Polarizability, represented by "AATS 5p", also emerged as a critical factor in predicting both bioconcentration and ecotoxicity. Furthermore, "VP-3", which captures paths traced by valence electrons across three connected non-hydrogen atoms, also exhibited high mean absolute SHAP values, further emphasizing its importance in the model's predictions. Finally, "SpMax5_Bhm" and "ATS3m", which reflect molecular structural complexity by incorporating topology and molecular weight, were essential for the model's predictive accuracy. This observation aligns with established ecotoxicological knowledge that compounds with greater molecular weight and structural complexity often exhibit higher toxicity, in part because larger molecules tend to be more lipophilic.[90] Increased molecular mass has also been shown to correlate with enhanced acute toxicity,[91] consistent with the baseline narcotic mode of action, which involves nonspecific and reversible disruption of membrane function.[92] The patterns described above are clearly illustrated in the boxplots of these molecular descriptors, highlighting the most impactful features across the two classes in the multispecies

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

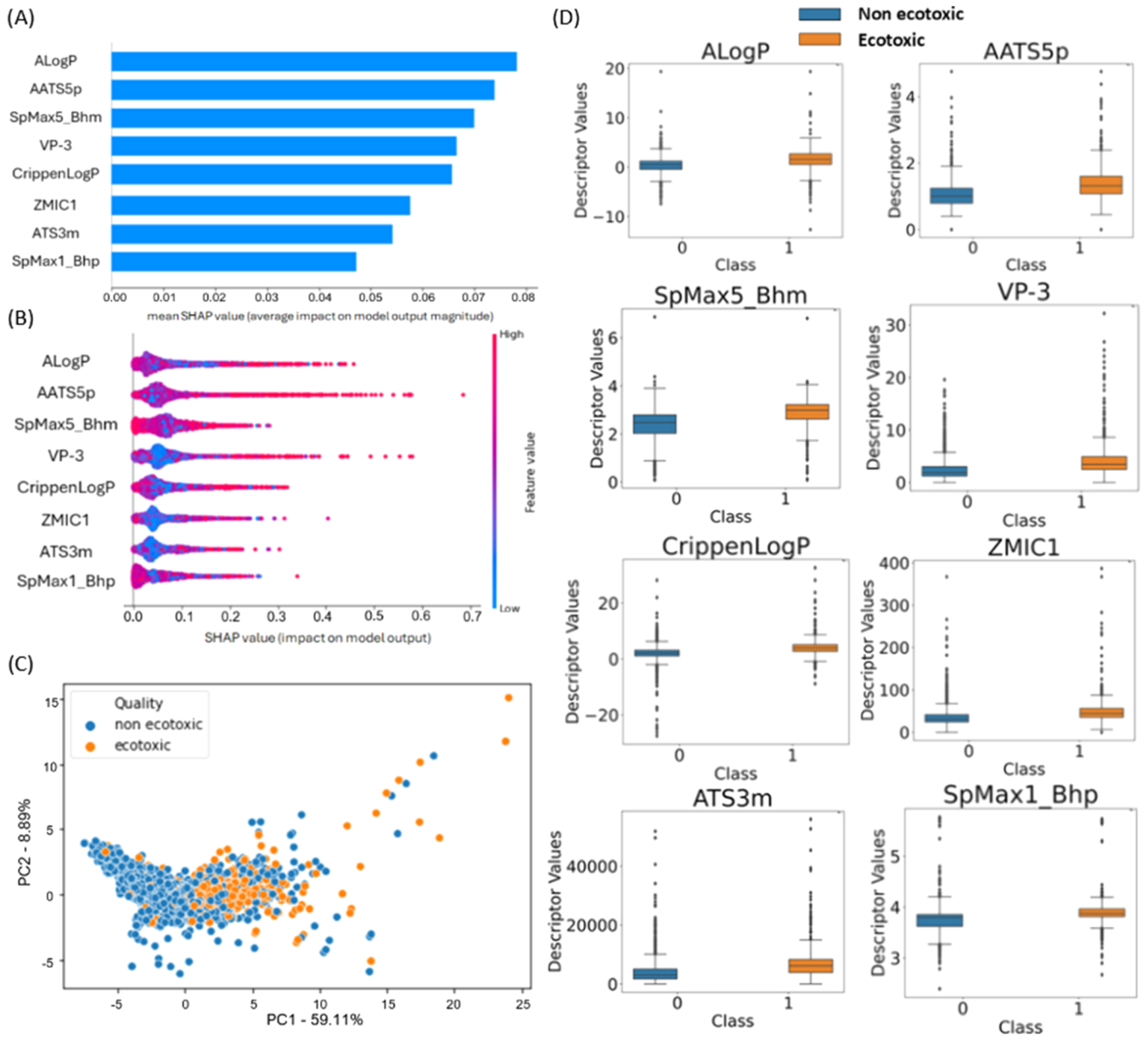


Figure 3. (A) Histogram of mean absolute SHAP values for the most influential features in differentiating ecotoxic compounds. (B) Distribution of local SHAP values for the most influential features in differentiating ecotoxic compounds. (C) 2D PCA plot depicting the training features of the ecotoxic ML model. (D) Boxplots representing the distribution values of the 8 most influential features in distinguishing ecotoxic compounds across the two ecotoxic classes.

fish data set (Figure 3). Our results therefore indicate that baseline narcosis is the dominant toxicity mechanism represented in our LC_{50} model. At the same time, descriptors related to molecular polarizability suggest that some compounds may also engage in more specific interactions (for example, through noncovalent binding to biomembranes or cellular receptors), indicating the potential presence of alternative modes of action in a subset of chemicals.

To complement the descriptor-based models, we also evaluated molecular fingerprint representations. As with the BCF analysis, the LC_{50} data set was modeled using Morgan fingerprints (various radii and bit lengths) and MACCS keys to capture structural features and support model performance. Table S12 presents the results of the most effective ML models after 10 epochs of external validation on randomly selected test sets. The ML models based on Morgan fingerprints demonstrated lower accuracy compared to the most effective descriptor-based model, which achieved a ROC AUC of 90.96 ± 1.35%. Among the Morgan-based models, the highest predictive performance was obtained using the Random Forest algorithm with Morgan fingerprints of radius 3 and 1024 bits as training features, yielding a ROC AUC of 79.31 ± 1.10%. Similarly, the MACCS-based ML model also exhibited lower predictive performance than the descriptor-based model, achieving a ROC AUC of 80.14 ± 1.70%. The SHAP analysis of the top-performing Morgan- and MACCS-based models, which aided in identifying structural alerts relevant for regulatory purposes, is presented in the SI and visualized in Figures S12−S15.

In the best MACCS-based ML model (which outperformed the other fingerprint-based models), bits 145 and 125, which represent the presence of single or fused aromatic rings, ranked

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Validation Metrics Obtained from Testing Publicly Available Online Software Using the t-F2F (Multiple Fish Species) and s-F2F-2 (*Pimephales promelas*) External Datasets of ADORE Database

online software/model	fish species	accuracy (%)	specificity (%)	sensitivity (%)	ROC AUC (%)
VEGA	multiple fish species	44.31	77.88	8.09	35.57
VEGA	*Pimephales promelas*	41.18	56.23	2.73	39.94
T.E.S.T	multiple fish species	67.18	82.05	51.14	65.15
T.E.S.T	*Pimephales promelas*	84.22	91.00	69.06	79.78
(our) G.AI.A MODEL	multiple fish species	91.12	96.61	84.80	97.00
(our) G.AI.A MODEL	*Pimephales promelas*	86.25	84.51	90.26	91.00

first and third in mean SHAP value (Figure S11A), respectively. These features were found in nearly 40% of ecotoxic compounds, compared to less than 20% in the nonecotoxic class. This aligns with previous studies suggesting that aromatic rings, such as carbonyl-benzene moieties,[93] contribute to high aquatic toxicity.[94−96] Additionally, bit 134, representing the presence of halogen atoms, was the most prevalent in the ecotoxic class (50%) and had the second-highest mean SHAP value. Similarly, bit 87, associated with halogen-containing fragments, also identified as an important feature in our best MACCS-based ML model for bioconcentration, was present in 40% of ecotoxic compounds but only in approximately 20% of nonecotoxic ones, though it had the lowest mean SHAP value. These findings are consistent with literature indicating that a higher number of halogen groups in chemical structures enhances aquatic toxicity.[97,98]

Comparison with Previous Studies for Ecotoxicity Classification

We conducted a comprehensive benchmarking study of our proposed ecotoxicity model, comparing its performance with other publicly available tools, including VEGA, and T.E.S.T. All models were evaluated using the same external data set of molecular compounds compiled for the development of our ML models (Table S13). Among the competing models, our G.AI.A model demonstrated superior performance with a ROC AUC of 90.96% outperforming T.E.S.T (ROC AUC = 67.00%) and VEGA (ROC AUC = 55.00). Accuracy and specificity metrics were comparable between our G.AI.A model (87.99 and 92.21%, respectively), VEGA (85.04 and 88.11%, respectively), and T.E.S.T (78.49 and 80.97%, respectively). Additionally, VEGA (33.33%) demonstrated very low sensitivity, highlighting its limited ability to differentiate ecotoxic compounds from nonecotoxic ones. In contrast, our G.AI.A model (73.38%) and T.E.S.T (66.90%) showed comparable effectiveness in identifying ecotoxic compounds, which represent the minority class in our data set. In addition, we benchmarked our model against the two software tools using the publicly available external data set from VEGA (VEGA *Fathead Minnow* LC_{50}), with the results summarized in Table S14. Our model exhibited comparable performance in accuracy, specificity, and ROC AUC, while achieving higher sensitivity (87.69%). T.E.S.T., which is trained exclusively on LC_{50} data for *Pimephales promelas*, showed notable improvements in accuracy, specificity, and ROC AUC but maintained relatively low sensitivity (63.38%). VEGA demonstrated a modest increase in specificity and ROC AUC but continued to show low sensitivity (33.78%). Overall, these results underscore the strong generalizability of our model relative to the other tools. Finally, while other studies and tools focus less on model interpretability and chemical insights, our G.AI.A model additionally provides valuable understanding of the chemical groups associated with ecotoxicity.

The ExtraTrees model demonstrates a strong overall accuracy of 87.99%; however, its sensitivity, which reflects its ability to correctly identify ecotoxic compounds, remains at an acceptable level of 73.38%. Despite this, the model exhibits robust generalization capabilities when evaluated on external data sets, as shown by validation using the benchmark data set ADORE. Two external subsets were selected for evaluation: t_F2F (1,118 compounds across multiple fish species) and s-F2F-2 (497 compounds specific to *Pimephales promelas*). Figure S16 illustrates the overlap and distinctions in chemical space between these external data sets and our LC_{50} training set showcasing the wide applicability domain of our data set compared to a well-established external LC_{50} data set. Comprehensive details on preprocessing steps, species representation, and data set complexity are provided in the SI. The results of the benchmarking study for the three models are summarized in Table 3. Regarding the t_F2F data set, our ExtraTrees model outperformed the others, achieving an accuracy of 90.44% and a specificity of 96.42%, compared to VEGA (44.31% accuracy, 77.88% specificity) and T.E.S.T (67.18% accuracy, 82.05% specificity). Additionally, our G.AI.A model demonstrated a significantly improved sensitivity of 83.88%, surpassing both the validation results within our data set and the sensitivity scores of VEGA (8.09%) and T.E.S.T (51.14%). Furthermore, our G.AI.A model exhibited strong robustness and stability, with a ROC AUC of 96.06%, outperforming VEGA (35.57%) and T.E.S.T (65.15%).

However, to account for the limitation of T.E.S.T, which was trained exclusively on LC_{50} data from the fish species *Pimephales promelas*, we also evaluated all models on a dedicated data set (s-F2F-2) containing experimental LC_{50} data specifically for this species. On this data set, our G.AI.A model achieved the highest ROC AUC (90.01%) and sensitivity (89.12%), while maintaining comparable accuracy to T.E.S.T (84.62% vs 84.22%). However, our G.AI.A model showed a slightly lower ability to distinguish nonecotoxic compounds, with a specificity of 82.59% compared to 91.00% for T.E.S.T. Therefore, our G.AI.A model demonstrates robustness and accuracy both when tested on specific fish species and when applied to external data sets that it has not encountered before.

While both models identify lipophilicity-related descriptors as influential, consistent with fundamental principles of environmental chemistry, they address different toxicological end points and capture distinct aspects of chemical risk. The BCF model characterizes the potential for chronic ecological risk by predicting long-term bioaccumulation in aquatic organisms. Its key descriptors (e.g., CrippenLogP, TopoPSA, AATS 5p) reflect properties governing bioconcentration and partitioning behavior, aligning with established mechanistic understanding. In contrast, the LC_{50} model targets acute toxicity, representing the immediate hazard posed by a

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

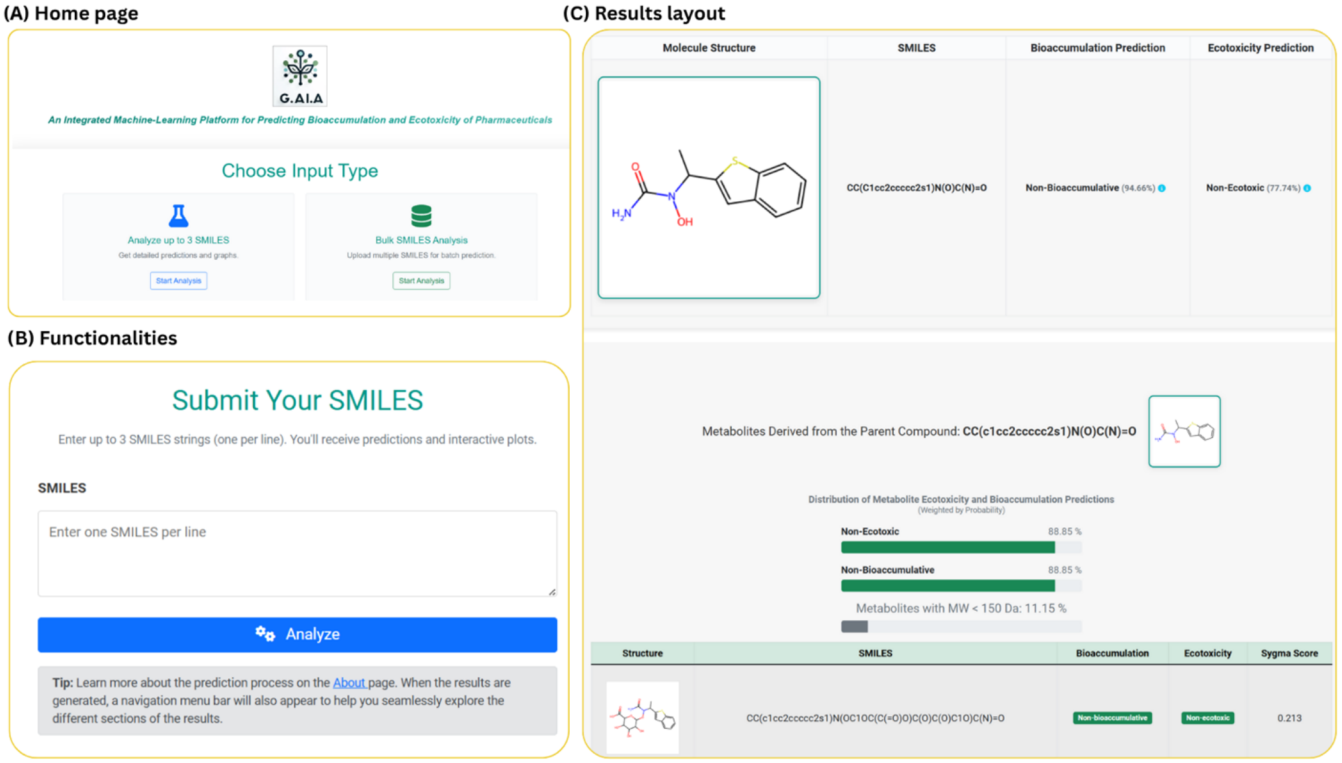


Figure 4. Screenshots from the G.AI.A platform. (A) Homepage view. (B) User interface displayed upon selecting the "Analyse up to 3 SMILES" mode. (C) Output summary following input of the SMILES for the antiasthmatic drug zileuton, showing predictions from the machine learning models for the parent compound and its metabolites.

compound at short-term exposure (up to 96 h). Although lipophilicity contributes to acute toxicity, the LC_{50} model additionally incorporates structural and electronic descriptors (e.g., SpMax5_Bhm, ATS3m, VP-3) associated with molecular size, electronic distribution, and toxicodynamic interactions such as membrane disruption or receptor binding processes, distinct from those driving bioaccumulation. Thus, the LC_{50} model is not redundant, rather, it complements the BCF model by capturing different mechanisms and exposure time scales: BCF relates to long-term accumulation and chronic effects, whereas LC_{50} reflects immediate toxic impact. Together, these end points provide a more complete and regulatory-relevant assessment of ecological risk.

Assessment of Ecotoxicity and Bioconcentration of Pharmaceutical Metabolites

To enhance the overall environmental safety profile of a drug molecule, it is crucial to assess the ecotoxicity of its metabolites, considering human metabolism and their subsequent excretion into the environment. Consequently, predicting a compound's metabolites is a key step in evaluating its full environmental impact. To this end, we evaluated the accuracy and robustness of three publicly available software tools, Metapredictor, SyGMa, Metatrans, for metabolite prediction using a validation set of 221 drugs from CHEMBL. The drugs were used as input for all three software, and the predicted metabolites were compared to the experimentally reported chemical structures available in ChEMBL. The accuracy was determined by calculating the ratio of identified experimental metabolites to the total predicted metabolites, resulting in 47% accuracy for Metapredictor, 46% accuracy for SyGMa, 38% for Metatrans.

Since all the software tools generated a large number of metabolic products for several drugs, an additional filtering step was introduced into the pipeline. We set the cutoff at 24 predicted metabolites per drug as a deliberate, empirical, data-informed choice. This number corresponds to the highest count of experimentally confirmed metabolites for any single compound in our reference data set (nicotine), ensuring we capture the full range of known metabolites (high sensitivity) while excluding unlikely low-probability candidates beyond this range (high specificity). At the same time, this threshold imposes a biologically informed upper bound that filters out exceedingly rare or implausible transformations. In essence, it captures all plausible metabolites reported for any drug to date, but constrained enough to avoid the combinatorial explosion of trivial predictions that often plagues metabolite prediction tools.[99] After applying this 24-metabolite filter, we recalculated the ratio of identified experimental metabolites to the total predicted metabolites for all drugs, which resulted in 47% accuracy for Metapredictor, 47% for SyGMa and 38% for Metatrans. Additionally, the mean ratio per drug was computed, yielding improved accuracy percentages for two out of three software 52% for SyGMa and 43% for Metatrans-, while for Metapredictor the mean ratio remained the same value 47%.

Due to its superior performance, SyGMa was selected for integration into our ecotoxicity and bioaccumulation pipeline. It is used to predict the metabolites of candidate drugs, providing with only the 24 most probable metabolites, when available. These predicted metabolites then undergo further evaluation for potential classification as ecotoxic and/or bioaccumulative.

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Webserver Implementation

The objective of this computational study was to develop innovative *in silico* tools to enhance screening processes in the early stages of drug discovery. To facilitate this, we developed an online tool, G.AI.A, which integrates the top-performing machine learning models: a Gradient Boosting model using PaDEL descriptors for bioconcentration prediction and an ExtraTrees model, also based on PaDEL descriptors, for ecotoxicity prediction. Additionally, the tool incorporates SyGMA software for drug metabolite prediction, as depicted in Figure 4. The application domain of G.AI.A lies in predictive toxicology, enabling researchers and regulatory bodies to assess the toxicological profiles of small organic compounds, excluding those containing heavy metals, by analyzing their chemical structures. It facilitates the identification of ecotoxic potential and supports green drug development through computational filtering and prioritization. The model's reliability is reinforced by previously achieved ROC AUC scores of 94.60% for bioconcentration and 96.06% for ecotoxicity, demonstrating its strong predictive performance. The platform was implemented using Django (back-end in Python, front-end in HTML/CSS/JavaScript). Users can submit SMILES strings of organic compounds through two modes offered by the Web server:

1. **"Analyse up to 3 SMILES" mode** - users can input up to three SMILES strings for analysis. The tool outputs bioconcentration and ecotoxicity classifications, visualizes molecular structures, provides classification confidence scores, and displays predicted metabolites along with their associated predicted classes. Metabolites with molecular weights below 150 Da are excluded from classification. Metabolite relevance is ranked using SyGMa scores, and the likelihood of class 0 (nontoxic, nonbioaccumulative) outcomes is presented as a weighted percentage based on predicted probabilities. Interactive t-SNE plots (Figure S17), generated using Plotly, show the positioning of input compounds relative to the training data: class 0 and class 1 compounds are marked with green and red circles, respectively, while user-submitted compounds are shown as blue stars. Separate visualizations are provided for bioconcentration and ecotoxicity predictions. For compounds classified as class 1, the parent molecule's structural features contributing to this classification are highlighted. Additional insights include boxplots of key PaDEL descriptor distributions from the training data set, showing where the input compounds fall in relation to each class.
2. **"Bulk SMILES Analysis" mode** - users can upload structure files in ″.txt", ".csv", or ".xlsx" format for high-throughput screening. The tool generates a CSV output that includes the input SMILES strings, predicted bioconcentration and ecotoxicity classes along with their associated classification confidence scores, up to 24 most probable metabolites per compound (where applicable), and a summary of any entries that could not be processed. For deeper interpretation and visualization, users are encouraged to utilize the "Analyse up to 3 SMILES" mode.

To proceed with the analysis, SMILES strings must be valid, meaning they are RDKit-compatible and represent molecules with a molecular weight exceeding 150 Da.

CONCLUSIONS

Our study focuses on the application and interpretation of ML techniques to predict the bioconcentration potential and acute fish toxicity of a broad range of organic chemicals. These predictive models serve as green toxicology tools, offering valuable metrics that support medicinal chemists in designing safer chemicals and complement *in silico* strategies such as virtual screening and binding free energy calculations. By integrating such tools into early stage research, our approach contributes to a more sustainable and holistic drug discovery process. We trained five different ML classifiers using experimentally derived logBCF and LC_{50} values compiled through an extensive literature review. These models were rigorously benchmarked against publicly available tools using both our curated data sets and independent, unbiased external data sets. Results showed strong performance and robustness, with both bioconcentration and ecotoxicity models achieving ROC AUC values above 90%. In addition to predictive accuracy, the models also yielded mechanistic insights by identifying key physicochemical and structural features linked to bioaccumulation. External validation further confirmed the models' generalizability, as they consistently matched or outperformed existing tools across various test sets. To address the often-overlooked environmental risks posed by drug metabolites, we also explored predictive modeling in this area. Following a comprehensive review of metabolite prediction tools, we compiled a reference data set of 221 drugs and their experimentally verified metabolites from the ChEMBL database. Among the tools assessed, SyGMA showed the best performance, correctly predicting 52% of known metabolites, highlighting its utility for downstream toxicity screening. To translate these findings into a practical resource, we developed G.AI.A (https://gaiatox.eu/), an automated pipeline that combines our top-performing models for bioconcentration, ecotoxicity, and metabolite prediction. This integrated tool provides a powerful framework for assessing the environmental footprint of pharmaceuticals. With the potential to be implemented as an early stage screening step or regulatory aid, it supports the development of safer, more environmentally conscious drugs while minimizing animal testing and promoting proactive risk assessment.

ASSOCIATED CONTENT

Data Availability Statement

All calculations were performed with freely available tools and codes. All novel methods were comprehensively described in the methods section, in relevant workflows and Supporting Information. The Web server is freely accessible at https://gaiatox.eu/ and the data sets are available as CSV files in the Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c02286.

Additional information and supporting figures and tables regarding the data generation and description, feature selection and models construction, evaluation of model performance, prediction of drugs' metabolites, model performance on bioaccumulation and ecotoxicity classification and comparison with previous studies for ecotoxicity classification (PDF)

logBCF data sets (CSV)

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

LC_{50} data sets (CSV)

AUTHOR INFORMATION

Corresponding Authors

Panagiotis Zoumpoulakis − *Cloudpharm PC, Athens 15125, Greece; Department of Food Science and Technology, University of West Attica, Egaleo 12243, Greece;* Email: pzoump@uniwa.gr

Minos-Timotheos Matsoukas − *Cloudpharm PC, Athens 15125, Greece; Department of Biomedical Engineering, University of West Attica, Egaleo 12243, Greece;* orcid.org/0000-0002-4642-8163; Email: mmatsoukas@uniwa.gr

Authors

Evangelos Tsoukas − *Cloudpharm PC, Athens 15125, Greece*

Michail Papadourakis − *Cloudpharm PC, Athens 15125, Greece*

Eleni Chontzopoulou − *Cloudpharm PC, Athens 15125, Greece*

Spyridon Vythoulkas − *Cloudpharm PC, Athens 15125, Greece*

Christos Didachos − *Cloudpharm PC, Athens 15125, Greece*

Dionisis Cavouras − *Department of Biomedical Engineering, University of West Attica, Egaleo 12243, Greece*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c02286

Author Contributions

[||]E.T. and M.P. contributed equally to this study. E.T., M.P., E.C., and D.C. model development, data curation, methodology and formal analysis, C.D.: data curation, E.T. and S.V. software, E.T., E.C., and M.P.: writing-original draft preparation. E.T., M.P., E.C., D.C., P.Z., and M.-T.M.: review and editing. D.C., P.Z., and M.-T.M: supervision. P.Z. and M.-T.M: conceptualization and resources. All authors have given approval to the final version of the manuscript.

Funding

The open access publishing of this article is financially supported by HEAL-Link.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Vasilis Panagiotopoulos, Sotiris Ouzounis, for useful technical-related suggestions, Marina Roussaki for resources management and Georgia Fragiadaki, Christos Lyssas and Nikolaos Fytilis for technical assistance. The authors thank the European Union's Horizon Europe Research and Innovation Programme for their funding facilitating this study through the Enviromed project (Grant agreement No. 101057844).

REFERENCES

(1) *European Green Deal - European Commission.* https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en (accessed June 6, 2024).

(2) Rampi, V.; Bisazza, O. The EU Green Deal: The Challenge of Greening Medical Technologies. *Clin. Chem. Lab. Med.* **2023**, *61* (4), 651−653.

(3) Nations, U. *Paris Agreement.* United Nations. https://www.un.org/en/climatechange/paris-agreement (accessed June 6, 2024).

(4) European Parliament and Council of European Union. Regulation (EC) No 1272/2008 on Classification, Labelling and Packaging of Substances and Mixtures, Amending and Repealing Directives 67/548/EEC and 1999/45/EC, and Amending Regulation (EC) No 1907/2006. *Off J. Eur. Union* **2008**, *50*, 1−1355.

(5) *European Commission 'REACH Revision under the Chemicals Strategy'.* https://environment.ec.europa.eu/topics/chemicals/reach-regulation_en (accessed June 6, 2024).

(6) *European Green Deal: Commission seeks views on reviewing rules on hazardous substances in electrical and electronic equipment - European Commission.* https://environment.ec.europa.eu/news/european-green-deal-commission-seeks-views-reviewing-rules-hazardous-substances-electrical-and-2022-03-10_en (accessed June 6, 2024).

(7) Ozben, T.; Fragão-Marques, M. Chemical Strategies for Sustainable Medical Laboratories. *Clin. Chem. Lab. Med.* **2023**, *61* (4), 642−650.

(8) Pereira, A. M. P. T.; Silva, L. J. G.; Lino, C. M.; Meisel, L. M.; Pena, A. A Critical Evaluation of Different Parameters for Estimating Pharmaceutical Exposure Seeking an Improved Environmental Risk Assessment. *Sci. Total Environ.* **2017**, *603*−*604*, 226−236.

(9) Białk-Bielińska, A.; Stolte, S.; Arning, J.; Uebers, U.; Böschen, A.; Stepnowski, P.; Matzke, M. Ecotoxicity Evaluation of Selected Sulfonamides. *Chemosphere* **2011**, *85* (6), 928−933.

(10) Grabarczyk, Ł.; Mulkiewicz, E.; Stolte, S.; Puckowski, A.; Pazda, M.; Stepnowski, P.; Białk-Bielińska, A. Ecotoxicity Screening Evaluation of Selected Pharmaceuticals and Their Transformation Products towards Various Organisms. *Environ. Sci. Pollut. Res.* **2020**, *27* (21), 26103−26114.

(11) Banjare, P.; Matore, B.; Singh, J.; Roy, P. P. In Silico Local QSAR Modeling of Bioconcentration Factor of Organophosphate Pesticides. *In Silico Pharmacol.* **2021**, *9* (1), No. 28.

(12) Lunghini, F.; Marcou, G.; Azam, P.; Patoux, R.; Enrici, M. H.; Bonachera, F.; Horvath, D.; Varnek, A. QSPR Models for Bioconcentration Factor (BCF): Are They Able to Predict Data of Industrial Interest? *SAR QSAR Environ. Res.* **2019**, *30* (7), 507−524.

(13) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput.-Aided Mol. Des.* **2011**, *25* (6), 533−554.

(14) Khan, P. M.; Baderna, D.; Lombardo, A.; Roy, K.; Benfenati, E. Chemometric Modeling to Predict Air Half-Life of Persistent Organic Pollutants (POPs). *J. Hazard. Mater.* **2020**, *382*, No. 121035.

(15) Khan, K.; Kumar, V.; Colombo, E.; Lombardo, A.; Benfenati, E.; Roy, K. Intelligent Consensus Predictions of Bioconcentration Factor of Pharmaceuticals Using 2D and Fragment-Based Descriptors. *Environ. Int.* **2022**, *170*, No. 107625.

(16) Kobayashi, Y.; Yoshida, K. Development of QSAR Models for Prediction of Fish Bioconcentration Factors Using Physicochemical Properties and Molecular Descriptors with Machine Learning Algorithms. *Ecol. Inf.* **2021**, *63*, No. 101285.

(17) Zhao, L.; Montanari, F.; Heberle, H.; Schmidt, S. Modeling Bioconcentration Factors in Fish with Explainable Deep Learning. *Artif. Intell. Life Sci.* **2022**, *2*, No. 100047.

(18) Xu, J.-Y.; Wang, K.; Men, S.-H.; Yang, Y.; Zhou, Q.; Yan, Z.-G. QSAR-QSIIR-Based Prediction of Bioconcentration Factor Using Machine Learning and Preliminary Application. *Environ. Int.* **2023**, *177*, No. 108003.

(19) Pore, S.; Pelloux, A.; Chatterjee, M.; Banerjee, A.; Roy, K. Machine Learning-Based q-RASAR Predictions of the Bioconcentration Factor of Organic Molecules Estimated Following the Organisation for Economic Co-Operation and Development Guideline 305. *J. Hazard. Mater.* **2024**, *479*, No. 135725.

(20) Chen, Z.; Li, N.; Li, L.; Liu, Z.; Zhao, W.; Li, Y.; Huang, X.; Li, X. BCDPi: An Interpretable Multitask Deep Neural Network Model for Predicting Chemical Bioconcentration in Fish. *Environ. Res.* **2025**, *264*, No. 120356.
(21) Hartung, T. ToxAIcology - The Evolving Role of Artificial Intelligence in Advancing Toxicology and Modernizing Regulatory Science. *ALTEX* **2023**, *40* (4), 559−570.
(22) Reuschenbach, P.; Silvani, M.; Dammann, M.; Warnecke, D.; Knacker, T. ECOSAR Model Performance with a Large Test Set of Industrial Chemicals. *Chemosphere* **2008**, *71* (10), 1986−1995.
(23) US EPA, O. *Toxicity Estimation Software Tool (TEST)*. https://www.epa.gov/comptox-tools/toxicity-estimation-software-tool-test (accessed Feb 27, 2025).
(24) Khan, K.; Baderna, D.; Cappelli, C.; Toma, C.; Lombardo, A.; Roy, K.; Benfenati, E. Ecotoxicological QSAR Modeling of Organic Compounds against Fish: Application of Fragment Based Descriptors in Feature Analysis. *Aquat. Toxicol.* **2019**, *212*, 162−174.
(25) Schlender, T.; Viljanen, M.; van Rijn, J. N.; Mohr, F.; Peijnenburg, W. JGM.; Hoos, H. H.; Rorije, E.; Wong, A. The Bigger Fish: A Comparison of Meta-Learning QSAR Models on Low-Resourced Aquatic Toxicity Regression Tasks. *Environ. Sci. Technol.* **2023**, *57* (46), 17818−17830.
(26) Zubrod, J. P.; Galic, N.; Vaugeois, M.; Dreier, D. A. Physiological Variables in Machine Learning QSARs Allow for Both Cross-Chemical and Cross-Species Predictions. *Ecotoxicol. Environ. Saf.* **2023**, *263*, No. 115250.
(27) He, Y.; Liu, G.; Hu, S.; Wang, X.; Jia, J.; Zhou, H.; Yan, X. Implementing Comprehensive Machine Learning Models of Multi-species Toxicity Assessment to Improve Regulation of Organic Compounds. *J. Hazard. Mater.* **2023**, *458*, No. 131942.
(28) Li, F.; Fan, D.; Wang, H.; Yang, H.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Pesticide Aquatic Toxicity with Chemical Category Approaches †Electronic Supplementary Information (ESI) Available: The SMILES Strings and Toxic Classes of All Chemicals Are Listed in SI1 of ESI, and the Performance of Binary and Ternary Classification Models for Ten-Fold Cross-Validation Using Different Fingerprints and Modeling Methods Is Listed in SI2. See DOI: 10.1039/C7tx00144d. *Toxicol. Res.* **2017**, *6* (6), 831−842.
(29) Tuulaikhuu, B.-A.; Guasch, H.; García-Berthou, E. Examining Predictors of Chemical Toxicity in Freshwater Fish Using the Random Forest Technique. *Environ. Sci. Pollut. Res.* **2017**, *24* (11), 10172−10181.
(30) Wu, J.; D'Ambrosi, S.; Ammann, L.; Stadnicka-Michalak, J.; Schirmer, K.; Baity-Jesi, M. Predicting Chemical Hazard across Taxa through Machine Learning. *Environ. Int.* **2022**, *163*, No. 107184.
(31) Xu, M.; Yang, H.; Liu, G.; Tang, Y.; Li, W. In Silico Prediction of Chemical Aquatic Toxicity by Multiple Machine Learning and Deep Learning Approaches. *J. Appl. Toxicol.* **2022**, *42* (11), 1766−1776.
(32) Viljanen, M.; Minnema, J.; Wassenaar, P. N. H.; Rorije, E.; Peijnenburg, W. What Is the Ecotoxicity of a given Chemical for a given Aquatic Species? Predicting Interactions between Species and Chemicals Using Recommender System Techniques. *SAR QSAR Environ. Res.* **2023**, *34* (10), 765−788.
(33) Schür, C.; Gasser, L.; Perez-Cruz, F.; Schirmer, K.; Baity-Jesi, M. A Benchmark Dataset for Machine Learning in Ecotoxicology. *Sci. Data* **2023**, *10* (1), 718.
(34) Gasser, L.; Schür, C.; Perez-Cruz, F.; Schirmer, K.; Baity-Jesi, M. Machine Learning-Based Prediction of Fish Acute Mortality: Implementation, Interpretation, and Regulatory Relevance. *Environ. Sci.: Adv.* **2024**, *3* (8), 1124−1138.
(35) Furuhama, A.; Toida, T.; Nishikawa, N.; Aoki, Y.; Yoshioka, Y.; Shiraishi, H. Development of an Ecotoxicity QSAR Model for the KAshinhou Tool for Ecotoxicity (KATE) System, March 2009 Version. *SAR QSAR Environ. Res.* **2010**, *21* (5−6), 403−413.
(36) Gramatica, P.; Cassani, S.; Sangion, A. Aquatic Ecotoxicity of Personal Care Products: QSAR Models and Ranking for Prioritization and Safer Alternatives' Design. *Green Chem.* **2016**, *18* (16), 4393−4406.
(37) Singh, K. P.; Gupta, S.; Kumar, A.; Mohan, D. Multispecies QSAR Modeling for Predicting the Aquatic Toxicity of Diverse Organic Chemicals for Regulatory Toxicology. *Chem. Res. Toxicol.* **2014**, *27* (5), 741−753.
(38) Sheffield, T. Y.; Judson, R. S. Ensemble QSAR Modeling to Predict Multispecies Fish Toxicity Lethal Concentrations and Points of Departure. *Environ. Sci. Technol.* **2019**, *53* (21), 12793−12802.
(39) Lane, T. R.; Harris, J.; Urbina, F.; Ekins, S. Comparing LD50/LC50 Machine Learning Models for Multiple Species. *ACS Chem. Health Saf.* **2023**, *30* (2), 83−97.
(40) Ai, H.; Wu, X.; Zhang, L.; Qi, M.; Zhao, Y.; Zhao, Q.; Zhao, J.; Liu, H. QSAR Modelling Study of the Bioconcentration Factor and Toxicity of Organic Compounds to Aquatic Organisms Using Machine Learning and Ensemble Methods. *Ecotoxicol. Environ. Saf.* **2019**, *179*, 71−78.
(41) Fliszkiewicz, B.; Sajdak, M. Fragments Quantum Descriptors in Classification of Bio-Accumulative Compounds. *J. Mol. Graphics Modell.* **2023**, *125*, No. 108584.
(42) Zhu, S.; Nguyen, B. H.; Xia, Y.; Frost, K.; Xie, S.; Viswanathan, V.; Smith, J. A. Improved Environmental Chemistry Property Prediction of Molecules with Graph Machine Learning. *Green Chem.* **2023**, *25* (17), 6612−6617.
(43) Ibáñez, M.; Bijlsma, L.; Pitarch, E.; López, F. J.; Hernández, F. Occurrence of Pharmaceutical Metabolites and Transformation Products in the Aquatic Environment of the Mediterranean Area. *Trends Environ. Anal. Chem.* **2021**, *29*, No. e00118.
(44) Celiz, M. D.; Tso, J.; Aga, D. S. Pharmaceutical Metabolites in the Environment: Analytical Challenges and Ecological Risks. *Environ. Toxicol. Chem.* **2010**, *28* (12), 2473−2484.
(45) Bottoni, P.; Caroli, S.; Caracciolo, A. B. Pharmaceuticals as Priority Water Contaminants. *Toxicol. Environ. Chem.* **2010**, *92* (3), 549−565.
(46) Wassenaar, P. N. H.; Verbruggen, E. M. J.; Cieraad, E.; Peijnenburg, W. J. G. M.; Vijver, M. G. Variability in Fish Bioconcentration Factors: Influences of Study Design and Consequences for Regulation. *Chemosphere* **2020**, *239*, No. 124731.
(47) Miller, T. H.; Gallidabino, M. D.; MacRae, J. I.; Owen, S. F.; Bury, N. R.; Barron, L. P. Prediction of Bioconcentration Factors in Fish and Invertebrates Using Machine Learning. *Sci. Total Environ.* **2019**, *648*, 80−89.
(48) Yuan, J.; Xie, C.; Zhang, T.; Sun, J.; Yuan, X.; Yu, S.; Zhang, Y.; Cao, Y.; Yu, X.; Yang, X.; Yao, W. Linear and Nonlinear Models for Predicting Fish Bioconcentration Factors for Pesticides. *Chemosphere* **2016**, *156*, 334−340.
(49) European Chemicals Agency. *Guidance on Information Requirements and Chemical Safety Assessment: Chapter R.11: PBT and vPvB Assessment.*; Publications Office: LU, 2017.
(50) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3* (1), No. 33.
(51) Xiao, Z.; Zhu, M.; Chen, J.; You, Z. Integrated Transfer Learning and Multitask Learning Strategies to Construct Graph Neural Network Models for Predicting Bioaccumulation Parameters of Chemicals. *Environ. Sci. Technol.* **2024**, *58* (35), 15650−15660.
(52) OECD. Test No. 203: Fish, Acute Toxicity Test. *OECD Guidelines for the Testing of Chemicals, Section 2* 2025. DOI: 10.1787/9789264069961-en.
(53) Connors, K. A.; Beasley, A.; Barron, M. G.; Belanger, S. E.; Bonnell, M.; Brill, J. L.; de Zwart, D.; Kienzler, A.; Krailler, J.; Otter, R.; Phillips, J. L.; Embry, M. R. Creation of a Curated Aquatic Toxicology Database: EnviroTox. *Environ. Toxicol. Chem.* **2019**, *38* (5), 1062−1073.
(54) QMRF_BCF_CAESAR.Pdf. https://www.vegahub.eu/vegahub-dwn/qmrf/QMRF_BCF_CAESAR.pdf (accessed Nov 18, 2025).
(55) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J. Cheminf.* **2018**, *10* (1), No. 10.

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(56) QMRF_FATHEAD_LC50_KNN.Pdf. https://www.vegahub.eu/vegahub-dwn/qmrf/QMRF_FATHEAD_LC50_KNN.pdf (accessed Nov 18, 2025).
(57) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−1474.
(58) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.
(59) 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* (2), 107−113.
(60) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; Gedeck, P.; Jones, G.; NadineSchneider; Kawashima, E.; Nealschneider, D.; Dalke, A.; Swain, M.; Cole, B.; Turk, S.; Savelev, A.; tadhurst-cdd; Vaucher, A.; Wójcikowski, M.; Take, I.; Scalfani, V. F.; Walker, R.; Ujihara, K.; Probst, D.; Lehtivarjo, J.; Faara, H.; godin guillaume; Pahl, A.; Monat, J. Rdkit/Rdkit: 2024_09_5 (Q3 2024) Release. 2025 DOI: 10.5281/zenodo.14779836.
(61) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(62) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63* (1), 3−42.
(63) 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; KDD '16*; Association for Computing Machinery: New York, NY, USA, 2016; pp 785−794 DOI: 10.1145/2939672.2939785.
(64) Freund, Y.; Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55* (1), 119−139.
(65) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5), 1189−1232.
(66) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.* **2002**, *16* (1), 321−357.
(67) Sasada, T.; Liu, Z.; Baba, T.; Hatano, K.; Kimura, Y. A Resampling Method for Imbalanced Datasets Considering Noise and Overlap. *Procedia Comput. Sci.* **2020**, *176*, 420−429.
(68) Momeny, M.; Neshat, A. A.; Hussain, M. A.; Kia, S.; Marhamati, M.; Jahanbakhshi, A.; Hamarneh, G. Learning-to-Augment Strategy Using Noisy and Denoised Data: Improving Generalizability of Deep CNN for the Detection of COVID-19 in X-Ray Images. *Comput. Biol. Med.* **2021**, *136*, No. 104704.
(69) Baratloo, A.; Hosseini, M.; Negida, A.; El Ashal, G. Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emerg (Tehran)* **2015**, *3* (2), 48−49.
(70) Mangalathu, S.; Hwang, S.-H.; Jeon, J.-S. Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based SHapley Additive exPlanations (SHAP) Approach. *Eng. Struct.* **2020**, *219*, No. 110927.
(71) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc., A* **2016**, *374* (2065), No. 20150202.
(72) Hart, A. Mann-Whitney Test Is Not Just a Test of Medians: Differences in Spread Can Be Important. *BMJ* **2001**, *323* (7309), 391−393.
(73) Ridder, L.; Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **2008**, *3* (5), 821−832.
(74) Zhu, K.; Huang, M.; Wang, Y.; Gu, Y.; Li, W.; Liu, G.; Tang, Y. MetaPredictor: In Silico Prediction of Drug Metabolites Based on Deep Language Models with Prompt Engineering. *Briefings Bioinf.* **2024**, *25* (5), No. bbae374.
(75) Litsa, E. E.; Das, P.; Kavraki, L. E. Prediction of Drug Metabolites Using Neural Machine Translation. *Chem. Sci.* **2020**, *11* (47), 12777−12788.
(76) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kizilören, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2024**, *52* (D1), D1180−D1192.
(77) Wishart, D. S.; Tian, S.; Allen, D.; Oler, E.; Peters, H.; Lui, V. W.; Gautam, V.; Djoumbou-Feunang, Y.; Greiner, R.; Metz, T. O. BioTransformer 3.0—a Web Server for Accurately Predicting Metabolic Transformation Products. *Nucleic Acids Res.* **2022**, *50* (W1), W115−W123.
(78) Wicker, J.; Lorsbach, T.; Gütlein, M.; Schmid, E.; Latino, D.; Kramer, S.; Fenner, K. enviPath − The Environmental Contaminant Biotransformation Pathway Resource. *Nucleic Acids Res.* **2016**, *44* (Database issue), D502−D508.
(79) Pramanik, S.; Roy, K. Modeling Bioconcentration Factor (BCF) Using Mechanistically Interpretable Descriptors Computed from Open Source Tool "PaDEL-Descriptor. *Environ. Sci. Pollut. Res.* **2014**, *21* (4), 2955−2965.
(80) Bertato, L.; Chirico, N.; Papa, E. Predicting the Bioconcentration Factor in Fish from Molecular Structures. *Toxics* **2022**, *10* (10), 581.
(81) Grisoni, F.; Consonni, V.; Villa, S.; Vighi, M.; Todeschini, R. QSAR Models for Bioconcentration: Is the Increase in the Complexity Justified by More Accurate Predictions? *Chemosphere* **2015**, *127*, 171−179.
(82) Papa, E.; Dearden, J. C.; Gramatica, P. Linear QSAR Regression Models for the Prediction of Bioconcentration Factors by Physicochemical Properties and Structural Theoretical Molecular Descriptors. *Chemosphere* **2007**, *67* (2), 351−358.
(83) Li, X.; Anderson, J. S. M.; Jobst, K. J. Bioaccumulative Chemicals Are Either Too Hard or Too Soft: Conceptual Density Functional Theory as a Screening Tool for Emerging Pollutants. *Environ. Int.* **2024**, *183*, No. 108388.
(84) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and Prediction of Soil Sorption Coefficients of Non-Ionic Organic Pesticides by Molecular Descriptors. *Chemosphere* **2000**, *41* (5), 763−777.
(85) Chiodi, D.; Ishihara, Y. Magic Chloro": Profound Effects of the Chlorine Atom in Drug Discovery. *J. Med. Chem.* **2023**, *66* (8), 5305−5331.
(86) Zhang, J.; Zhang, X.; Hu, T.; Xu, X.; Zhao, D.; Wang, X.; Li, L.; Yuan, X.; Song, C.; Zhao, S. Polycyclic Aromatic Hydrocarbons (PAHs) and Antibiotics in Oil-Contaminated Aquaculture Areas: Bioaccumulation, Influencing Factors, and Human Health Risks. *J. Hazard. Mater.* **2022**, *437*, No. 129365.
(87) Naumann, K. Influence of Chlorine Substituents on Biological Activity of Chemicals: A Review. *Pest Manage. Sci.* **2000**, *56* (1), 3−21.
(88) Valsecchi, C.; Grisoni, F.; Consonni, V.; Ballabio, D. Structural Alerts for the Identification of Bioaccumulative Compounds. *Integr. Environ. Assess. Manage.* **2019**, *15* (1), 19−28.
(89) Benfenati, E.; Manganaro, A.; Gini, G. VEGA.*QSAR: AI inside a Platform for Predictive Toxicology*.
(90) Newman, M. C.; Newman, M. C. *Fundamentals of Ecotoxicology: The Science of Pollution*, 4th ed.; CRC Press: Boca Raton, 2014. DOI: 10.1201/b17658.
(91) Delistraty, D. Acute Toxicity to Rats and Trout with a Focus on Inhalation and Aquatic Exposures. *Ecotoxicol. Environ. Saf.* **2000**, *46* (2), 225−233.
(92) Klüver, N.; Vogs, C.; Altenburger, R.; Escher, B. I.; Scholz, S. Development of a General Baseline Toxicity QSAR Model for the Fish Embryo Acute Toxicity Test. *Chemosphere* **2016**, *164*, 164−173.
(93) Peets, P.; Wang, W.-C.; MacLeod, M.; Breitholtz, M.; Martin, J. W.; Kruve, A. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. *Environ. Sci. Technol.* **2022**, *56* (22), 15508−15517.
(94) Kumar, A.; Ojha, P. K.; Roy, K. Safer and Greener Chemicals for the Aquatic Ecosystem: Chemometric Modeling of the Prolonged

and Chronic Aquatic Toxicity of Chemicals on *Oryzias Latipes*. *Aquat. Toxicol.* **2024**, *273*, No. 106985.
(95) Khan, K.; Khan, P. M.; Lavado, G.; Valsecchi, C.; Pasqualini, J.; Baderna, D.; Marzo, M.; Lombardo, A.; Roy, K.; Benfenati, E. QSAR Modeling of *Daphnia Magna* and Fish Toxicities of Biocides Using 2D Descriptors. *Chemosphere* **2019**, *229*, 8−17.
(96) Zhang, R.; Guo, H.; Hua, Y.; Cui, X.; Shi, Y.; Li, X. Modeling and Insights into the Structural Basis of Chemical Acute Aquatic Toxicity. *Ecotoxicol. Environ. Saf.* **2022**, *242*, No. 113940.
(97) Papa, E.; van der Wal, L.; Arnot, J. A.; Gramatica, P. Metabolic Biotransformation Half-Lives in Fish: QSAR Modeling and Consensus Analysis. *Sci. Total Environ.* **2014**, *470*−*471*, 1040−1046.
(98) Li, F.; Sun, G.; Fan, T.; Zhang, N.; Zhao, L.; Zhong, R.; Peng, Y. Ecotoxicological QSAR Modelling of the Acute Toxicity of Fused and Non-Fused Polycyclic Aromatic Hydrocarbons (FNFPAHs) against Two Aquatic Organisms: Consensus Modelling and Comparison with ECOSAR. *Aquat. Toxicol.* **2023**, *255*, No. 106393.
(99) de Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism. *Front. Chem.* **2019**, *7*, No. 402, DOI: 10.3389/fchem.2019.00402.

https://doi.org/10.1021/acs.jcim.5c02286
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX