New
Contaminants

https://doi.org/10.48130/newcontam-0026-0010

Original Research

Open Access

Toxicity prediction and ecological risk assessment of new contaminants to rare and endangered species using machine learning-QSAR: a case study of conserving *Gobiocypris rarus* in the Yangtze River Basin

Ying Wang[1*], Xin Wang[1,2], Yunchi Zhou[1], Yinghao Cheng[3], Xiaomin Li[1], Xiaolei Wang[4], Yuefei Ruan[5,6], Zhaomin Dong[1,7] and Wenhong Fan[1*]

Received: 16 January 2026

Revised: 24 March 2026

Accepted: 10 April 2026

Published online: 30 April 2026

Abstract

Rare and endangered species play a crucial role in ecosystems, but pollutants, especially new contaminants such as per- and polyfluoroalkyl substances (PFASs), increasingly threaten their survival. It is vital to conduct ecological risk assessments for these species. However, conventional toxicity testing methods are ethically and practically infeasible for rare and endangered species. Using modeling approaches to predict toxicity is a viable solution to assess ecological risks. This study proposed a machine learning, quantitative structure-activity relationship (ML-QSAR) method for predicting toxicity, using a case study of *Gobiocypris rarus*, an endangered fish species endemic to China's Yangtze River Basin. Six algorithms were applied to construct models based on molecular descriptors and life stage encoding. The random forest-based model showed optimal performance for predicting both acute and chronic toxicity (R^2_{Acute} = 0.99, $R^2_{Chronic}$ = 0.93). Feature importance analysis revealed a key mechanistic divergence: life stage was the dominant factor in acute toxicity; in contrast, chronic toxicity was primarily driven by molecular interaction descriptors. The optimal models were applied to predict the toxicity of 73 pollutants detected in the *Gobiocypris rarus*' habitat. The ecological risk assessment indicated that current environmental concentrations of 12 PFASs remain low, but it is important to note their persistence and potential long-term ecological risks. This study presents a novel, non-testing approach for quantifying chemical threats to endangered species for which data are limited. Integrating life-stage-specific mechanisms into an ML-QSAR method provides a powerful tool for supporting evidence-based conservation and pollutant management in vulnerable aquatic ecosystems.

Keywords: *Gobiocypris rarus*, Toxicity prediction, Machine learning, Quantitative structure-activity relationship (QSAR), Life stage, PFAS, Risk assessment

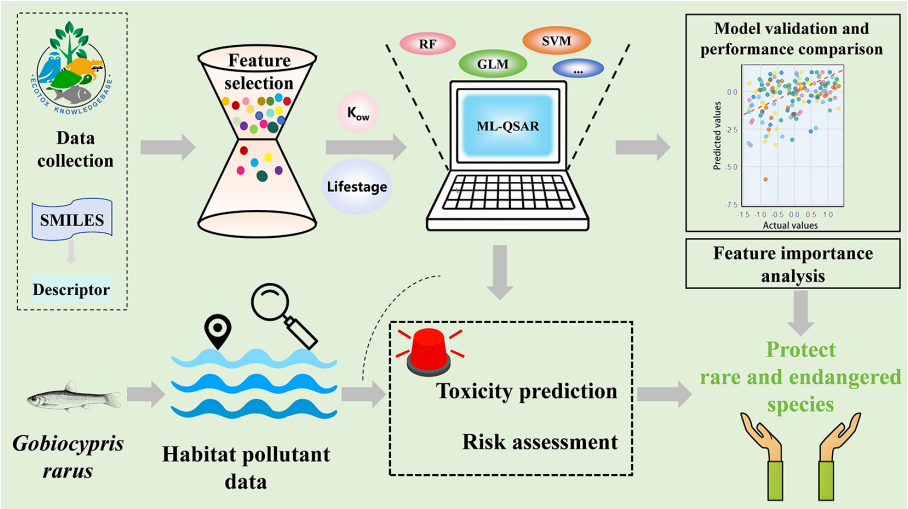
Highlights

- ML-QSAR models were developed to predict toxicity for the endangered fish *Gobiocypris rarus*.
- Life stage was the dominant factor in acute toxicity.
- Molecular interaction descriptors drove chronic toxicity predictions.
- The risk of new contaminants in the *Gobiocypris rarus*' habitat was assessed.
- Current PFAS levels pose low risk in the habitat, but long-term ecological concerns remain.

* Correspondence: Ying Wang (yingw@buaa.edu.cn or wy2012bnu@126.com); Wenhong Fan (fanwh@buaa.edu.cn)
Full list of author information is available at the end of the article.

https://doi.org/10.48130/newcontam-0026-0010

Graphical abstract



Introduction

Global aquatic ecosystems are increasingly threatened, particularly in major rivers and their tributaries. As a result, the survival of rare and endangered species is of growing concern[1]. These species play critical roles in ecosystems, and are focal points for biodiversity conservation due to their unique biological and ecological characteristics. Human activities, such as water pollution, overfishing, river channel modification, and habitat destruction, have led to sharp declines in the populations of numerous rare and endangered species. Some even face extinction in the wild[2]. Among these threats, pollution from a wide array of chemicals, especially new contaminants, creates significant challenges for the survival of rare and endangered species[3].

The Yangtze River Basin in China is one of the world's most iconic rivers and supports a rich array of aquatic biological resources[4], including many rare and endangered species such as *Acipenser sinensis*, *Myxocyprinus asiaticus*, and *Gobiocypris rarus* (*G. rarus*)[5]. These species play a crucial role in maintaining the ecological balance of the Yangtze River Basin[6]. The rapid pace of industrialization and urbanization has led to the emergence of new contaminants, such as per- and polyfluoroalkyl substances (PFASs) and conazoles, which are characterized by high environmental persistence and bioaccumulation potential. These new contaminants pose a significant threat to the aquatic ecosystems and rare and endangered species of the basin[7,8].

G. rarus is a small cyprinid fish endemic to the Yangtze River Basin[9]. It was designated a rare and endangered species in 1989 due to its restricted habitat range and sparse population[10]. Its high sensitivity to environmental fluctuations has made it particularly vulnerable to increased chemical pollution within the Yangtze River Basin, impacting its conservation status over time[11]. However, past toxicity studies have mainly focused on more widely distributed fish species in the basin, such as carp (*Cyprinus carpio*)[12] and catfish (*Tachysurus fulvidraco*)[13]. Less attention has been given to *G. rarus*, particularly regarding new contaminants. As such, they remain largely uncharacterized, hindering a systematic assessment of the ecological risks to its habitat. This lack of data for new pollutants in rare species like *G. rarus* represents a critical knowledge gap that this study aims to address.

Conducting biotoxicity experiments on rare and endangered species for a wide range of pollutants is impractical due to ethical considerations and the limited availability of artificially bred populations. This makes mathematical models a viable alternative for predicting toxicity. Among these approaches, quantitative structure-activity relationship (QSAR) models analyze the relationship between a compound's molecular structure and its biological activity. They have been widely applied to predict the toxicity of pollutants (including organic compounds and heavy metals)[14]. For example, Mu et al.[15] successfully used a QSAR approach to predict the acute toxicity of 25 metals or metalloids to aquatic organisms across five phyla and eight families.

Conventional QSAR models typically rely on low-dimensional molecular descriptors and linear regression methods[16], capturing limited structure-activity relationships and having a poor ability to address complex molecular structures or diverse biotoxicity responses. Consequently, for new pollutants with intricate molecular structures, or for rare and endangered species where there is little experimental data, conventional QSAR models do not adequately capture complex, multidimensional descriptors or establish nonlinear relationships[17]. This significantly reduces the applicability and accuracy of toxicity prediction models[18].

In recent years, machine learning (ML) approaches have achieved increasing success in developing QSAR models. Unlike conventional linear models, ML methods leverage high-dimensional molecular descriptors and address nonlinear relationships, significantly increasing prediction accuracy and model applicability[19]. These models autonomously identify and learn complex relationships between chemical structures and biological activity as a result of training on data. This gives them superior generalization ability in predicting the toxicity of novel chemical substances[20]. For example, Schmidt et al.[21] used an ML-enhanced QSAR model to successfully predict the toxicity of novel antibiotics to rainbow trout (*Oncorhynchus mykiss*), illustrating the potential of these models in ecotoxicological studies involving new contaminants and rare and endangered species (i.e., *G. rarus*).

This study compiles and organizes toxicity data for *G. rarus*, screens common physicochemical property parameters of pollutants, and calculates simplified molecular input line entry system (SMILES)-based molecular descriptors. Building on this foundation,

six ML algorithms (including random forest, support vector machine, and neural network) are used to construct ML-QSAR toxicity prediction models for pollutants affecting *G. rarus*. Following model validation, the ML-QSAR model with the highest predictive performance is selected to estimate the toxicity of key pollutants in *G. rarus*' habitat and assess associated ecological risks. Furthermore, the differences in acute and chronic toxicity mechanisms of related pollutants to *G. rarus* are discussed based on the results of feature importance analysis. The results enable the rapid and accurate prediction of biotoxicity for this rare and endangered species in the Yangtze River Basin. This offers a novel approach to obtaining toxicity data for such species. It provides a scientific basis for conserving *G. rarus* in the Yangtze River Basin and managing new pollutants.

Methods and materials

Datasets

The dependent variable was the toxicity data of *G. rarus*. It was sourced from the U.S. Environmental Protection Agency's (EPA) ECOTOX database (https://cfpub.epa.gov/ecotox/) using keywords such as '*G. rarus*', 'rare gudgeon', 'Chinese rare minnow', and '*Gobiocypris rarus*'. The search period ended in December 2024. The accuracy and reliability of the experimental data met standard methodological requirements. The specific criteria for screening toxicity data are as follows. (1) Distilled or deionized water must not be used as the experimental medium. (2) Experiments require a blank control group, in which test organisms exhibit no significant abnormal mortality or illness. (3) For acute fish toxicity tests, exposure durations must range between 1 and 4 d; chronic toxicity tests require exposures exceeding 21 d. (4) Toxicity data lacking simultaneous information on toxicity values, life stages, exposure durations, and test endpoints are excluded. (5) For acute toxicity test data, experimental species are not fed during the experiment to ensure data integrity and accuracy. Lethality was selected as the acute toxicity endpoint, including the median lethal concentration 50% (LC_{50}) and median effect concentration (EC_{50}). No observed effect concentration (NOEC), and lowest observed effect concentration (LOEC) were selected as the chronic toxicity endpoints[22].

It is important to note that toxicity data for new contaminants (e.g., PFASs and conazoles) were not obtained through a separate screening process. Rather, they were part of the complete dataset retrieved from ECOTOX using the species keywords described above, and they were subjected to the same screening criteria. New contaminants were subsequently identified from this complete dataset by cross-referencing compound names with established lists from recent literature. However, recognizing that ECOTOX may not fully cover the most recent toxicity data for new contaminants, we supplemented the ECOTOX-derived data with a targeted literature search in Web of Science and CNKI using specific new contaminant names combined with '*G. rarus*' and its synonyms. Data from this supplementary search were also screened using the same criteria to ensure consistency and quality.

The independent variables included parameters across multiple dimensions, including molecular descriptors (including physiochemical properties, topological features, and geometric information), the *n*-octanol-water partition coefficient (K_{ow}), and the life stage classification of *G. rarus*. The K_{ow} reflects the distribution and solubility characteristics of chemicals in aquatic environments; these variables significantly influence the metabolism and toxicity in aquatic organisms. Organic compounds with higher K_{ow} are more likely to accumulate in organisms, with stronger toxic effects[23].

The life stage of the organism is another significant factor that may affect toxicity. This stems from the substantial differences in sensitivity to pollutants between adult and juvenile fish. Juvenile fish have metabolic systems that are not yet fully developed and are generally more sensitive to pollutants. Adult fish may be more tolerant due to increased detoxification capabilities[24]. For example, fipronil ($C_{13}H_{11}Cl_2F_4N_3O$, log K_{ow} = 4.57) is toxic to *G. rarus* with LC_{50} values of 4.13 mg/L for juveniles, and 6.69 mg/L for adults[25]. As such, we classified and numerically encoded the life stages of *G. rarus*, including stages of embryo, juvenile, and adult, and incorporated 'life stage' as another key feature in the model (Supplementary Table S1). The benefits of this approach include: (1) numerical encoding (1, 2, 3) preserves the natural developmental order of life stages (embryo→juvenile→adult), which is biologically meaningful; (2) tree-based algorithms such as random forest handle numerical categorical variables effectively by splitting at thresholds; (3) this encoding method allows 'life stage' to be assessed as a single integrated feature in importance evaluation, thereby improving model interpretability.

Molecular descriptor generation

Generating molecular descriptors from the molecular structure of chemical substances is a crucial step in building QSAR models[26]. There is significant diversity in the naming of chemical substances, making it necessary to first convert the names of the substances into a unique representation of the compound's structure to enable the retrieval of molecular descriptors. In this study, SMILES strings were used to uniquely represent each chemical substance. SMILES is a common linear notation for entering and representing molecular structures; it provides a one-to-one correspondence with the molecular structure[27]. The notation is widely used to represent molecular structures and as an index to retrieve specific substances in chemical databases. The SMILES strings for all compounds in the cleaned and integrated dataset were retrieved using their names and stored using PubChempy in Python, enabling batch retrieval of SMILES strings by compound names[28]. Based on the SMILES strings of each compound, the molecular descriptors such as physicochemical properties, topological features, and geometric information were calculated using the Mordred package in Python[29,30]. The generated descriptors of each compound were then applied in subsequent ML-QSAR modeling without additional standardization. For metal ions that do not have SMILES strings, molecular descriptors generated by Mordred and log Kow values could not be calculated. These missing values were coded as NA (not available) in the dataset. The Random Forest algorithm used in this study has built-in mechanisms for handling missing data through proximity-based imputation and surrogate splits, allowing these compounds to be included in model training without requiring complete descriptor sets. For organometallic compounds with defined molecular structures and valid SMILES strings, all descriptors were calculated normally. An example Python script for retrieving compound information (names, CAS numbers, and SMILES strings) from PubChem is provided in Supplementary Text S1.

Data preprocessing

SMILES strings typically generate an extremely large number of molecular descriptors. When this count exceeds the number of data samples, it increases the training cost of the model. Additionally, strong correlations between certain descriptors can lead to multicollinearity, causing model overfitting and reducing training effectiveness[31]. Therefore, feature selection was used to reduce dimensionality and increase feature information. It is worth noting that outlier removal was not performed to preserve all available data for this rare species.

https://doi.org/10.48130/newcontam-0026-0010

This process involved two steps.

(1) Correlation screening was conducted between the descriptors and the target variable (toxicity values for *G. rarus*) to eliminate irrelevant descriptors. Specifically, the Pearson correlation coefficient was calculated for each descriptor with respect to the toxicity values (Eq. (1)). A Pearson correlation coefficient greater than 0.4 generally indicates moderate correlation[32]. Therefore, descriptors with an absolute correlation coefficient less than 0.4 were removed, as they were considered redundant or irrelevant.

$$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}((x_i-\overline{x}))^2(y_i-\overline{y})^2}} \quad (1)$$

where, x_i and y_i represent the values of the descriptor and toxicity data for the i^{th} sample, respectively; $\overline{x}$ and $\overline{y}$ are their respective means; and n is the number of samples.

(2) Correlations were calculated between molecular descriptor variables, and highly correlated features were removed. Specifically, a greater than 0.8 absolute value of the correlation coefficient between two features indicated a strong linear relationship between them[33]. One of the two was therefore removed. The feature with greater explanatory power or higher information content was typically retained, and the one with greater redundancy was removed.

ML-QSAR model

Based on our previous study[34], six commonly used ML algorithms were employed for modeling: random forest (RF), generalized linear model (GLM), support vector machine (SVM), k-nearest neighbors (KNN), neural network (NNET), and generalized linear model boosting (GLMENT).

RF makes predictions by integrating multiple decision trees, effectively handling high-dimensional data and improving model accuracy[35]. In this study, the RF model optimized the *mtry* parameter (range: 1–10) using a grid search, where *mtry* determined the number of randomly selected variables for each split. The GLM is a flexible regression model that adapts to various data types and provides clear interpretations[36]. Here, the GLM utilized standard linear regression without hyperparameter optimization. An SVM classifies or regresses by identifying the optimal hyperplane, making it suitable for high-dimensional data and able to address complex nonlinear problems[37]. In this study, the SVM adopted the svmRadialSigma method to automatically optimize the kernel width (*sigma*) and penalty coefficient (*C*). A KNN makes predictions by calculating distances between samples, offering a simple and intuitive approach that is effective for multi-classification problems[38]. Here, the KNN algorithm automatically selected the optimal number of neighbors (*k*) through cross-validation. A NNET simulates neuronal connections in the human brain, enabling it to learn complex nonlinear relationships with robust fitting capabilities[39]. In this study, the NNET model used a single-hidden-layer architecture. The number of neurons (1−10) was optimized, while excluding the second and third hidden layers. The elastic net regression automatically tuned the L1/L2 mixing ratio (alpha) and regularization strength (lambda). GLMENT integrates generalized linear models with ensemble learning techniques, increasing predictive performance. It is effective for complex regression and classification challenges[40].

To account for the adaptability of these algorithms to diverse data types, the dataset was divided into training and test sets at a 4:1 ratio. The test set is an independent dataset for external validation that was not used to construct the model to ensure that the model did not learn from these data. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The R^2 indicated the model's capacity to explain data variability. RMSE and MAE quantified the discrepancy between predicted and actual values.

Model validation

The model's reliability was evaluated using a combination of internal and external validation approaches, ensuring robust generalization ability and stability.

Internal validation was performed using k-fold cross-validation, a technique commonly used to estimate performance and prevent overfitting. This study used a 10-fold cross-validation[41]. The model's robustness was assessed using the cross-validation correlation coefficient Q^2, RMSE, and MAE, as defined in Eqs (2)–(4).

$$Q^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n} \quad (4)$$

External validation was done using an independent test set, with the coefficient of determination (R_{ex}^2), as defined in Eq. (5). This measured the model's external predictive ability.

$$R_{ex}^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \quad (5)$$

where, y_i represents the true value of the i^{th} sample in external validation; $\hat{y}_i$ represents the predicted value of the i^{th} sample; $\overline{y}$ represents the mean of all true values, and n represents the number of samples. When $R^2 > 0.6$ and the differences between R^2 and Q^2 were less than 0.3, the model was considered reliable and performed effectively[42]. All algorithms were implemented using R software v4.3.1 for data preprocessing and model building.

In addition to internal and external validation, the applicability domain (AD) of the optimal random forest models was defined using the descriptor range approach to ensure reliable predictions for new chemicals[43]. This method defines the AD as the p-dimensional hyperrectangle bounded by the minimum and maximum values of each molecular descriptor in the training set. A new chemical is considered to be within the AD if, for all its descriptors used in the model, its descriptor values fall within the corresponding training set ranges. For the acute toxicity model and chronic toxicity model, the minimum and maximum values for each descriptor in the training sets were determined. The descriptor range approach was selected because it is simple, transparent, conservative, and widely used in QSAR studies[44]. For the pollutants detected in *G. rarus*' habitat, molecular descriptors were calculated and compared against the training set ranges. All compounds had descriptor values within the training set ranges for both acute and chronic models, confirming that they were within the AD and that the toxicity predictions are reliable.

Feature importance analysis

The best-performing model was further analyzed for feature importance. The feature importance was quantified by calculating the mean decrease in accuracy (MDA), defined as the average increase in the model's MAE when the values of a specific feature were randomly shuffled[43]. If the model's MAE significantly increased after shuffling a

particular feature, it indicated that the feature significantly contributed to the model's predictions. If the MAE changed little, it indicated a low feature contribution. MDA results were used to rank all features from highest to lowest importance to identify the key features that most impacted the model's predictions.

Toxicity prediction and risk assessment

Surveys were used to identify the habitat locations of *G. rarus* (Supplementary Fig. S1). Relevant studies were collected from Web of Science and China National Knowledge Infrastructure (CNKI), and data were compiled about the types of pollutants and their exposure concentrations in *G. rarus*' habitat. For the identified pollutants, molecular descriptors were calculated using SMILES codes. The optimal ML-QSAR model was then used to predict the toxicity values of each pollutant at different life stages of *G. rarus* and for different effect types (i.e., acute and chronic toxicity). Then, a risk assessment based on the quotient method was conducted. Generally, this method compares the actual monitored or model-estimated environmental exposure concentration with toxicity data characterizing the hazard level of the substance to calculate the risk quotient (RQ) (Eq. (6))[44]. This approach was suitable for assessing the toxicological effects of individual compounds in this study.

$$RQ = \frac{EEC}{Toxicity\ value} \quad (6)$$

where, *EEC* refers to the pollutant's actual exposure concentration in the habitat of *G. rarus*. The *Toxicity value* represents the predicted acute or chronic toxicity data of the pollutant to *G. rarus* based on the ML-QSAR in this study. A larger RQ value indicates greater risk. An RQ value below 1 indicates a relatively safe risk level.

Results and discussion

Description of toxicity data

During data screening, entries were excluded if they were missing key information, including life stage, effect, and endpoint concentration. Ultimately, 49 acute toxicity data points and 67 chronic toxicity data points were obtained, encompassing ten major chemical classes (Fig. 1, Supplementary Tables S2, S3). Both acute and chronic datasets were dominated by conazoles and per- and polyfluoroalkyl substances (PFASs), with 23 and 27 entries, respectively. The high environmental stability and bioaccumulative potential of both conazoles and PFASs often result in their increased concentrations in aquatic systems and ecosystems. This raises significant concerns regarding their toxicity and ecological risks to aquatic organisms, including *G. rarus*.

Conazoles are a common class of fungicides and are widely used to control plant diseases[45]. They include various active ingredients, such as triazoles, imidazoles, and sulfur ether oxides, showing broad-spectrum fungicidal activity[46]. These are typically high-efficacy fungicides, and even trace residues can be detected in the

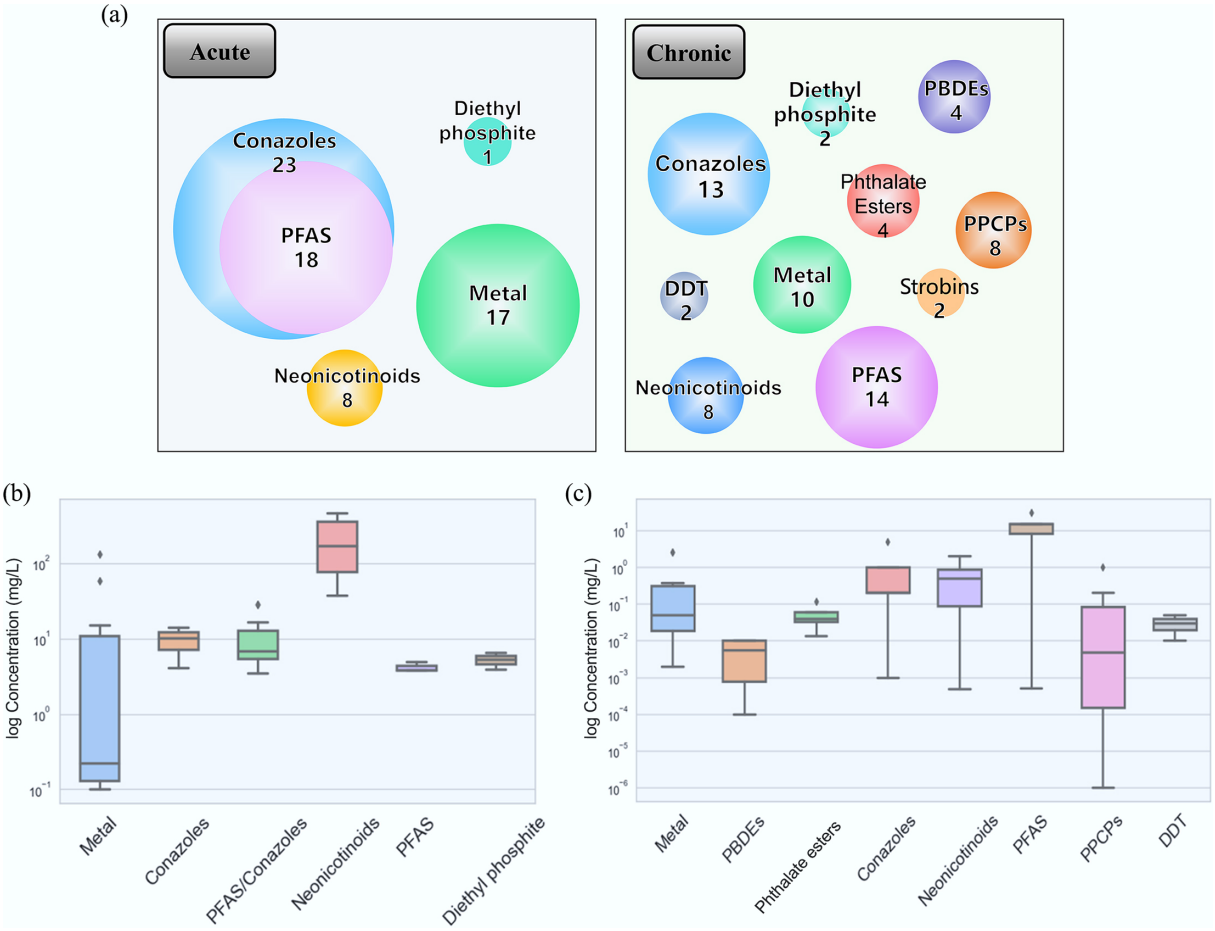


Fig. 1 Overview of the modeling data set. (**a**) Pollutant classifications in acute and chronic toxicity data sets. (**b**) Boxplot of acute toxicity for different pollutants. (**c**) Boxplot of chronic toxicity for different pollutants.

https://doi.org/10.48130/newcontam-0026-0010

environment[47]. Moreover, certain conazole fungicides may have prolonged residual periods, with extended environmental persistence and a higher likelihood of detection in environmental samples[48]. Conazoles have been found to have elevated acute toxicity to *G. rarus*. For example, tetraconazole ($C_{13}H_{11}Cl_2F_4N_3O$) exhibited a low LC_{50} value of 4.13 mg/L in juveniles, indicating particularly pronounced toxicity at this life stage[25].

PFASs form a class of chemicals with unique properties. They are extensively used in industrial and consumer products, including water-repellent, oil-resistant, and stain-resistant coatings[49]. Given their exceptional chemical stability and resistance to high temperatures, PFASs are highly persistent in the environment and resist degradation; they are referred to as 'forever chemicals'[50]. These properties facilitate their accumulation in natural environments, with potentially long-term impacts on water bodies, soils, and ecosystems[51].

In the chronic toxicity dataset, the fewest data points were identified for dichlorodiphenyltrichloroethane (DDT) and strontium bromate fungicides, with only two entries each. DDT is an organochlorine pesticide that has been banned for decades. However, it may still be in certain samples as a result of historical releases, due to its long-term environmental persistence[52]. As time passes, DDT concentrations have gradually decreased, leading to reduced concern regarding its toxicity. Strontium bromates fungicides are less commonly used, but still may be detected in specific environments due to their unique chemical properties. For example, decabromodiphenyl ether (deca-BDE, $C_{12}Br_{10}O$) has a Lowest Observed Effect Concentration (LOEC) as low as 0.01 mg/L for juvenile *G. rarus*, indicating elevated toxicity at early life stages. This compound has been linked to a mortality rate of 33% in *G. rarus*[53]. Future research should prioritize describing the toxicity mechanisms of these substances on *G. rarus* and strengthen their environmental management.

Feature selection

A total of 1,826 molecular descriptors were generated for each identified compound using the *Mordred* package in Python. During data preprocessing, correlation screening was performed separately on the acute and chronic toxicity datasets. Descriptors with minimal influence on the target variable were excluded according to the correlation between each descriptor and the toxicity values. The screening results showed that 92 descriptors were associated with acute toxicity; 125 descriptors were associated with chronic toxicity. The molecular descriptors selected to predict acute toxicity mainly relate to electronic properties, such as molecular orbital energy levels. It may be because acute toxicity typically depends on the chemical reactivity of substances and their short-term accumulation in organisms. In contrast, chronic toxicity primarily addresses the long-term effects of pollutants on organisms under prolonged exposure. This encompasses factors such as accumulation, metabolism, and genetic damage. Therefore, the descriptors retained to predict chronic toxicity are more closely associated with molecular polarity, lipophilicity, and stability. An additional correlation screening among the descriptors was then conducted, eliminating those with high collinearity. Ultimately, the acute toxicity dataset retained 14 features, and the chronic toxicity dataset retained 22 features (Supplementary Table S4). The descriptor selection process revealed that the dataset primarily comprises organic substances and spans various organism life stages. This is why, as discussed previously, the two feature parameters K_{ow} and life stage were incorporated in addition to the screened molecular descriptors.

ML-QSAR modeling

Based on the screened features and toxicity values, ML-QSAR models were developed to predict the acute and chronic toxicity of various pollutants to *G. rarus*. Internal and external validations found that the RF model outperformed the five other assessed models in predicting both acute and chronic toxicity (R^2_{Acute} = 0.99, $RMSE_{Acute}$ = 0.02, $R^2_{Chronic}$ = 0.93, $RMSE_{Chronic}$ = 1.40) (Figs 2a, 3a). These results demonstrated an excellent fit and robustness for the model. This finding aligns with the optimal model identified by Zhou et al.[54] in their study using ML to predict the adverse effects of metal nanomaterials on multiple aquatic organisms. The result showed that RF enhances overall predictive performance and mitigates overfitting by integrating the outputs of multiple decision trees when processing high-dimensional, nonlinear data.

When predicting acute toxicity, the GLM and SVM models performed at a moderate level (Fig. 2b, c). The GLM model faced difficulties with complex datasets due to its linear assumptions. The SVM model had advantages in handling high-dimensional data, but its efficacy was limited by the selection of kernel functions and parameter optimization. The KNN model performed relatively well (R_{ex}^2 = 0.75, RMSE = 1.61) (Fig. 2d), indicating a potential advantage when using small datasets. The NNET model performed poorly (R_{ex}^2 = 0.20, RMSE = 1.18) (Fig. 2e), likely due to its dependence on large datasets. Given the relatively small dataset in this study, the NNET model tended to memorize noise or specific details from the training data rather than capturing generalizable patterns. This led to suboptimal performance[55].

When predicting chronic toxicity, all algorithms achieved high coefficients of determination ($Q^2 > 0.7$, $R_{ex}^2 > 0.8$) (Fig. 3), indicating robust performance across models. This improved outcome may be attributed to the larger dataset available for predicting chronic toxicity.

Feature importance analysis and related toxicity mechanisms

The feature importance ranking results of the acute ML-QSAR model indicate that CIC_3, life stage, and ATSC2v significantly influenced model performance (Fig. 4a). CIC_3 quantifies the complementary information content of specific molecular segments. In the context of PFAS toxicity, the high importance of CIC3 likely reflects the specific binding affinity of PFAS molecules with proteins such as fatty acid-binding proteins (FABPs). Recent studies have shown that PFAS disrupt lipid metabolism and transport by binding to FABPs, leading to developmental toxicity in fish[56]. This molecular interaction mechanism is directly captured by the CIC3 descriptor, which quantifies the complementarity between pollutant molecules and biological targets[54]. Fish toxicity typically involves the interaction of molecules with receptors or enzymes within the organism. CIC_3 enables the quantification of specific aspects of these interactions. This provides deeper insights into how molecules engage with biological systems (Fig. 4c)[56,57]. This complex molecular descriptor also increases the ability to generalize to unknown chemical substances, improving the accuracy of toxicity predictions. Specifically, the CIC_3 descriptor describes the interaction mechanisms between pollutant molecules and target receptors in fish. For example, certain PFASs or conazoles can trigger toxic responses by binding to receptors in fish organisms[58]. By quantifying the complementarity and interactions between these molecules, CIC_3 helps determine the toxic effects of specific pollutants on *G. rarus* in this study.

As introduced previously, *G. rarus* may have different physiological, behavioral, and environmental adaptation capabilities at different life stages, influencing their response to potentially toxic

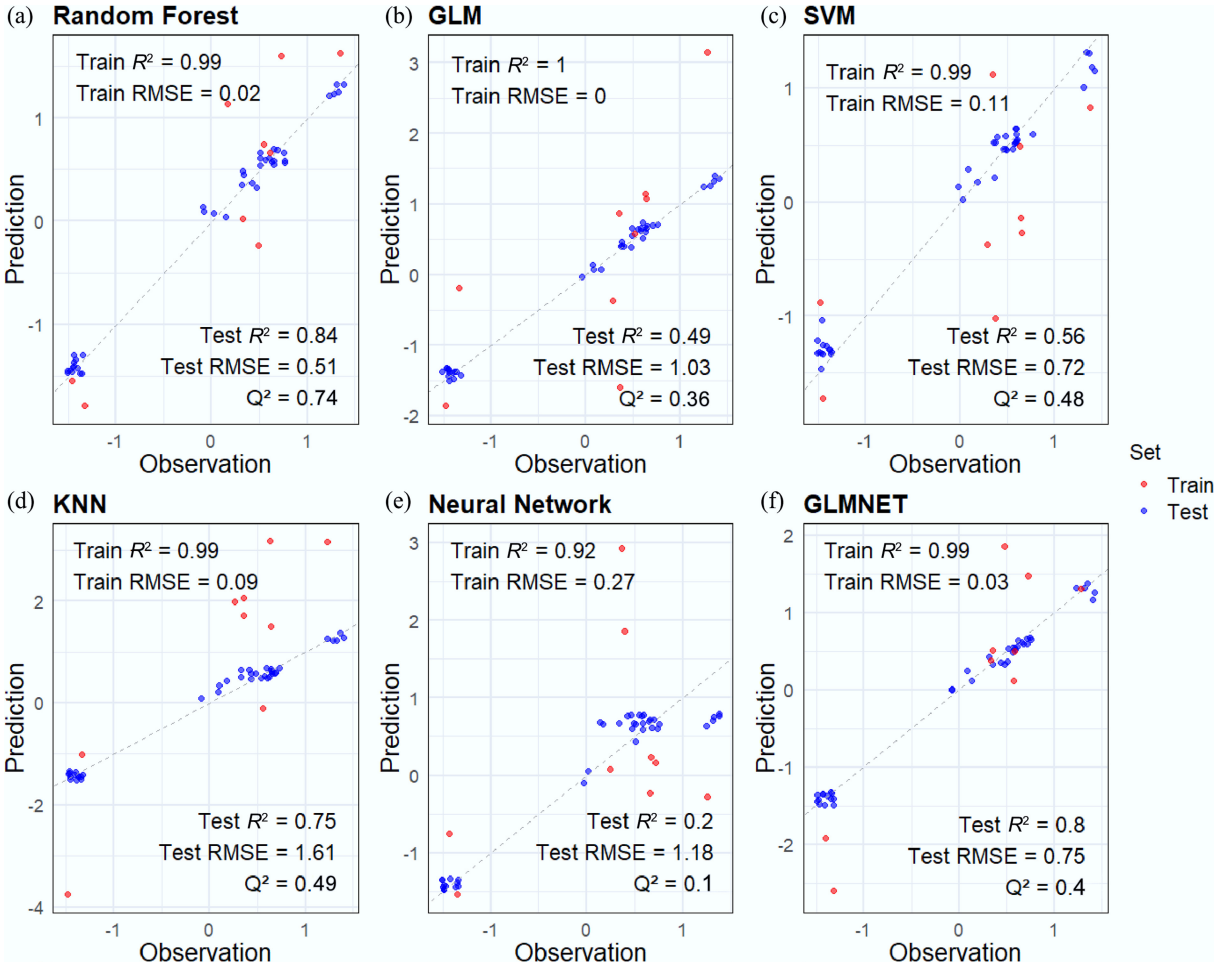


Fig. 2 Acute ML-QSAR model performance of different algorithms: (**a**) RF, (**b**) GLM, (**c**) SVM, (**d**) KNN, (**e**) Neural Network,and (**f**) GLMNET. The data points for the training set are plotted as blue dots. The data points for the test set are plotted as red dots. Train R^2 represents R^2, and test R^2 represents R_{ex}^2. R^2 and R_{ex}^2 values are rounded to two decimal places.

substances[59]. Comparing acute toxicity data revealed that *G. rarus* is less tolerant and more sensitive to most substances during the embryonic and larval stages. For example, the acute toxicity values for bisphenol A (BPA) were 16.57 mg/L for embryos, 16.45 mg/L for larvae, and 23.13 mg/L for adult fish. This may be because the embryonic and larval stages are critical developmental periods, and physiological functions, metabolic pathways, and detoxification capacities are weaker than later in life. Adult fish may have more robust metabolic and excretory capacities, somewhat mitigating the toxic effects of substances[60]. However, for PFASs, adult fish exhibit higher toxicity sensitivity compared to embryos and larvae, indicating there may be substance-specific metabolic pathways in the adult stage. This apparent contradiction can also be explained by toxicokinetic principles. PFASs are known to bind strongly to proteins, particularly serum albumin and fatty acid-binding proteins. Adult fish have higher protein concentrations and more developed enterohepatic circulation, which can lead to longer retention and accumulation of PFASs[58]. Additionally, adult fish may have specific transport proteins that facilitate PFAS uptake in certain organs, such as the liver and kidney, resulting in higher target organ concentrations despite lower overall uptake rates. These substance-specific toxicokinetic pathways highlight the complexity of life-stage-dependent toxicity and underscore the importance of incorporating life stage information in predictive models. In addition, fish at different life stages may show distinct behavioral patterns and ecological niches, which could influence their frequency of exposure and contact levels with toxic substances[61].

The ATSC2v descriptor is also a critical factor in the acute ML-QSAR model. This descriptor quantifies spatial arrangement and interactions between atoms or groups within a molecule. It is highly sensitive to subtle variations in internal molecular structure (Fig. 4c), enabling ATSC2v to capture molecular structural features closely linked to fish toxicity. As noted above, fish toxicity often arises from interactions between molecules and biological receptors. These interactions are frequently influenced by minute structural details. For example, pollutants with distinctive spatial configurations[57], such as PFASs, can bind to target sites in fish through specific spatial relationships among atoms or groups, eliciting toxic responses. By capturing the spatial arrangement and relative positional relationships of atoms or groups within a molecule, the ATSC2v descriptor helps explain how molecules interact with targets in fish[62].

For the chronic ML-QSAR model, molecular descriptors such as ATSC2i and ATSC1p had a greater influence than life stage (Fig. 4b). The ATSC2i descriptor had a particularly strong effect, largely due to its integration of the Moreau-Broto autocorrelation function with ionization potential[57]. Ionization potential measures the energy required for an atom within a molecule to lose an electron and is closely tied to the molecule's chemical reactivity and activity[63].

https://doi.org/10.48130/newcontam-0026-0010

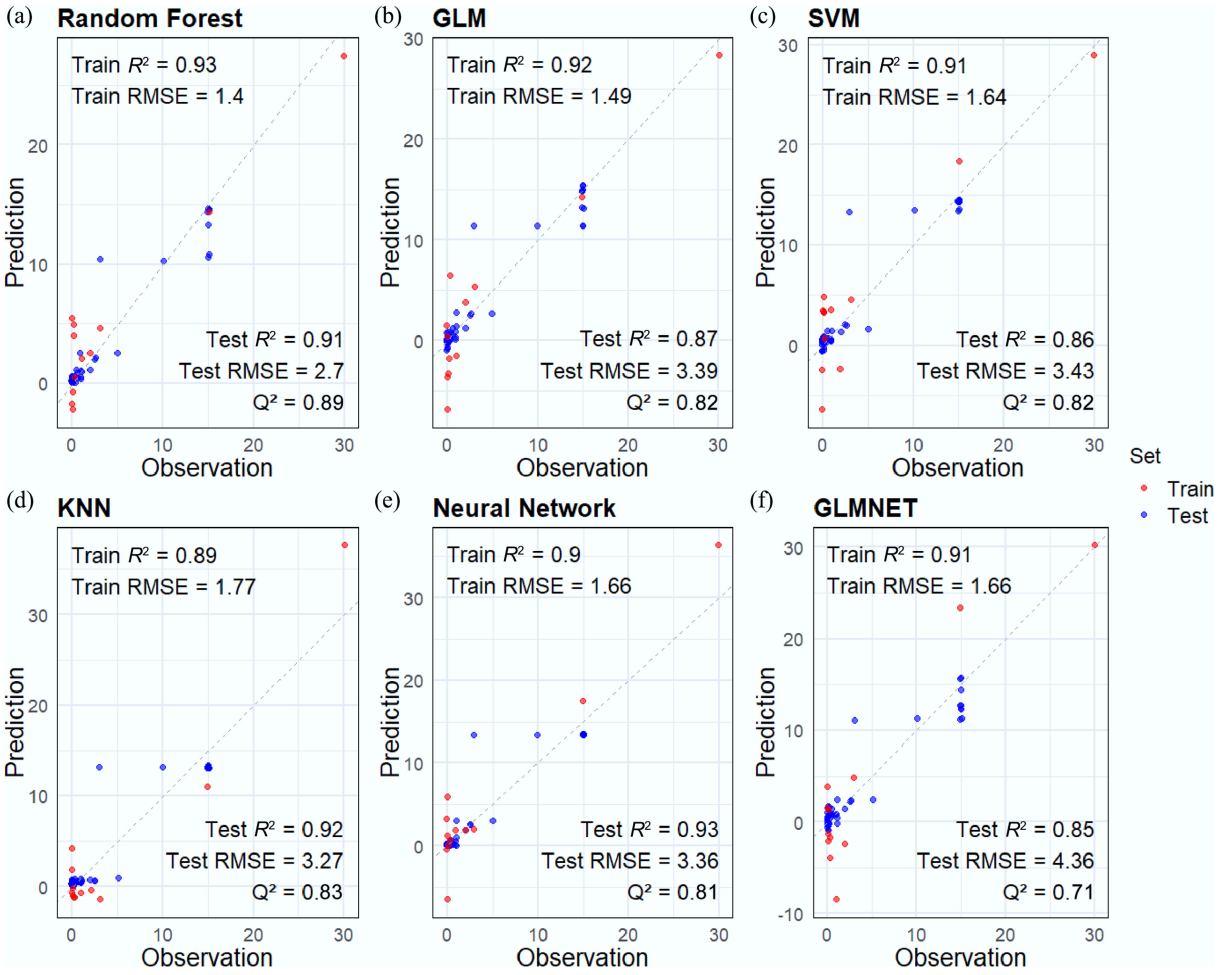


Fig. 3 Chronic ML-QSAR model performance of different algorithms: (**a**) RF, (**b**) GLM, (**c**) SVM, (**d**) KNN, (**e**) Neural Network, and (**f**) GLMNET. The data points for the training set are plotted as blue dots. The data points for the test set are plotted as red dots. Train R^2 represents R^2, and test R^2 represents R_{ex}^2. R^2 and R_{ex}^2 values are rounded to two decimal places.

Pollutant molecules with low ionization potential may be more susceptible to electron transfer reactions with receptors in fish, initiating toxic responses. The ATSC2i descriptor captures structural features related to ionization capability by weighting the autocorrelation function with ionization potential. This offers valuable information for predicting pollutant toxicity (Fig. 4c).

The ATSC1p descriptor integrates the Moreau-Broto autocorrelation function with polarizability[57]. Polarizability is a physical quantity that describes a molecule's ability to deform under an external electric field. This property is closely related to the molecule's electronic structure and reactivity (Fig. 4c)[57]. The ATSC1p descriptor emphasizes the relative positions and interactions between atoms or groups within the molecule by weighing the autocorrelation function with polarizability. Of particular interest are the correlations between atoms or groups separated by one position. The properties of these atoms or groups determine the intermolecular forces, such as van der Waals forces and hydrogen bonds. The toxicity of pollutants is often closely linked to their solubility and distribution characteristics within organisms, which are influenced by polarizability. For example, pollutant molecules with strong polar groups tend to have higher polarizability. This helps them form hydrogen bonds with water molecules, increasing their solubility in water. This increased solubility facilitates pollutant diffusion in aquatic environments, affecting the toxic responses in fish or other aquatic organisms[57]. Further, ATSC1p captures the relative positions and interactions of atoms or groups through the weighted autocorrelation function. Also, it describes how these structural features influence the molecule's environmental distribution and its interactions with biological targets.

In the context of chronic toxicity, life stage did not have a pronounced influence (mean concentration difference < 5%), likely due to the combined effects of many factors. First, differences in physiological and developmental stages may be critical in distinguishing between short-term and long-term effects. Acute toxicity typically focuses on the impact of high-dose exposures over a short period. Organisms may show heightened sensitivity during a specific studied period, particularly for juveniles or those in developmental stages. In contrast, chronic toxicity examines the effects of long-term, low-dose exposures, where organisms may have adapted mechanisms to mitigate the impact of toxic substances[64]. Second, variations in exposure duration and dosage may lead the factor of life stage to be more significant in acute toxicity prediction models, because in the short-term, high-dose exposures may have a greater effect on sensitive developmental stages[65]. Additionally, disparities in sample sizes across datasets may have contributed to the differing importance of life stage. In the acute toxicity dataset, the sample

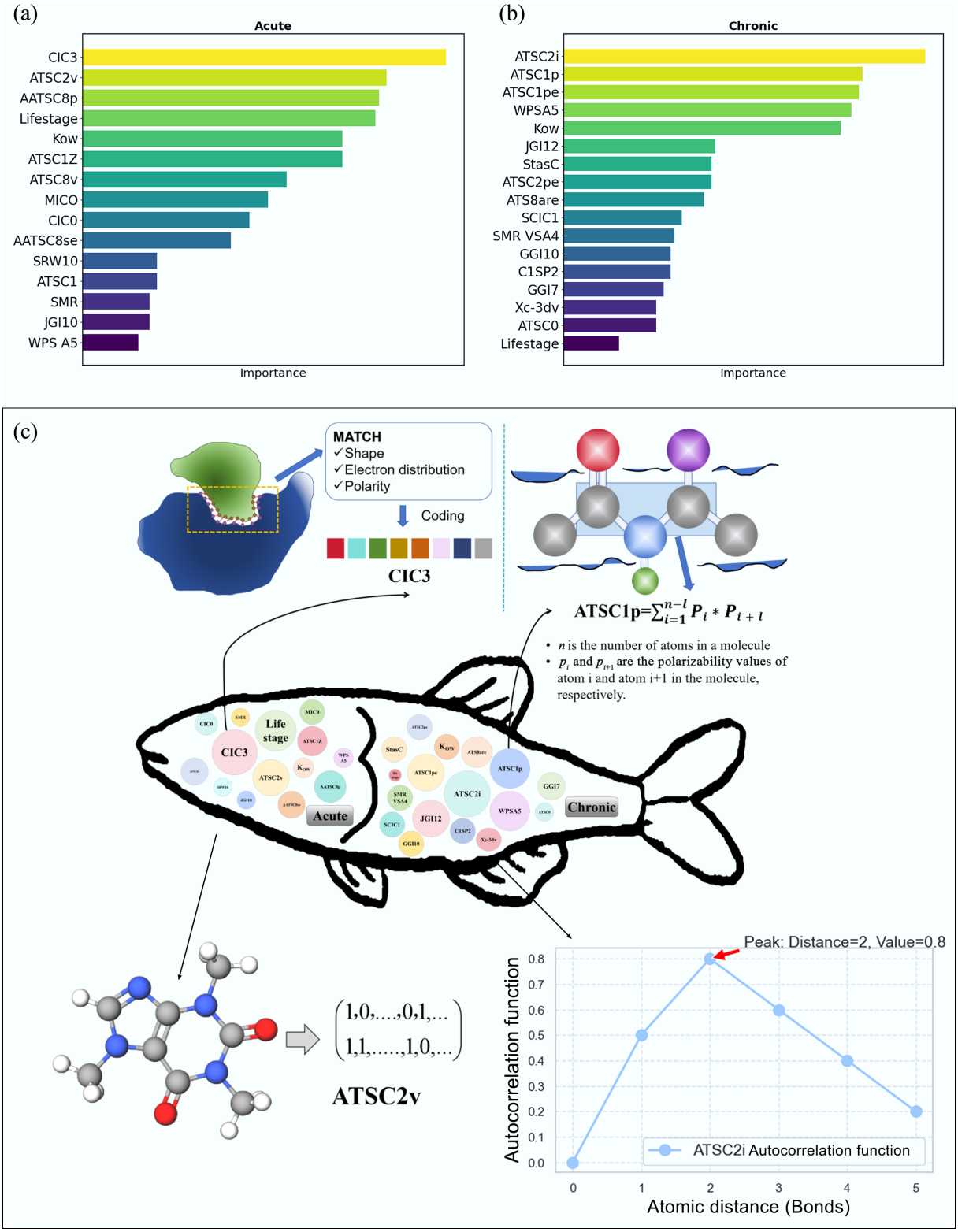


Fig. 4 Feature importance analysis of the model. Importance rank of different descriptors for ML-QSAR models: (**a**) acute, (**b**) chronic, and (**c**) descriptor explanations for CIC3, ATSC1P, ATSC2v, and ATSC2i.

sizes for embryos, juveniles, and adults were relatively balanced, clearly reflecting life stage effects. In contrast, in the chronic toxicity dataset, there were notably fewer samples for juveniles compared to adults. This likely reduced the observed impact of life stage.

It is worth noting that the *n*-octanol-water partition coefficient (K_{ow}) is a relatively important parameter characterizing the lipophilicity of compounds and has strong predictive importance in both acute and chronic toxicity models for *G. rarus*. Acute toxicity

https://doi.org/10.48130/newcontam-0026-0010

data show that high-K_{ow} substances (log K_{ow} > 5, e.g., $C_{14}H_{30}$, log K_{ow} ≈ 6.5) had significantly higher LC_{50} values at life stage_4 (215.66 mg/L) compared to low-K_{ow} substances (log K_{ow} < 3, e.g., C_7H_5NO, log K_{ow} ≈ 1.8, LC_{50} = 36.70 mg/L). This is consistent with the 'baseline toxicity' theory: lipophilic compounds more readily penetrate the biomembrane for accumulation; however, they require higher concentrations to trigger acute effects[66]. Chronic toxicity data showed that moderately high-K_{ow} compounds (3 < log K_{ow} < 5, e.g., PFOS, log K_{ow} ≈ 4.1) showed notable bioaccumulation potential under prolonged exposure (concentrations across life stage_1–4 stabilized at 1.57–1.59 mg/L). Low-K_{ow} substances (e.g., BPA, log K_{ow} ≈ 3) show < 2% variation in toxicity values across life stages due to rapid metabolism. The interaction between K_{ow} and life stages was particularly pronounced. In acute models, high-K_{ow} substances showed greater toxicity increases in the larval stage (life stage_2; e.g., $C_{16}H_{34}$, + 1.5%) compared to low-K_{ow} compounds (e.g., C_6H_7N, + 0.9%). This difference diminished in the chronic model results (e.g., $C_{16}H_{34}$ variations < 0.5% across life stages). This likely reflects metabolic adaptation partially offsetting K_{ow}'s initial influence during chronic exposure. This difference in acute compared to chronic models underscores K_{ow}'s dual role as an indicator of persistence/bioaccumulation; it indicates initial bioavailability in acute exposure, while also illustrating long-term kinetic processes in chronic scenarios. This study demonstrates that K_{ow} should be prioritized as a weighted feature in cross-life-stage toxicity prediction models.

Ecological risk assessment of the habitat of *G. rarus*

The toxicity of 73 pollutants reported in the habitat of *G. rarus* in the literature was predicted by using the optimal acute and chronic ML-QSAR models (Supplementary Tables S5, S6). Twelve PFASs pose a greater acute hazard to organisms over a short period (Supplementary Table S5). The pronounced acute toxicity of PFASs may be due to their ability to rapidly penetrate cell membranes and disrupt critical intracellular physiological processes[67]. These substances can induce acute toxic effects on organisms by impairing cellular respiration, altering membrane permeability, or causing DNA damage. In contrast, nonylphenol showed strong chronic toxicity, with a predicted value of 1.20 mg/L, which was below the overall average of 3.32 mg/L (Supplementary Table S6). The chronic toxicity shown for nonylphenol may be more complex, involving multifaceted effects such as immune system suppression, endocrine disruption, and organ damage under prolonged exposure[68]. In addition, PFASs may be more readily absorbed and metabolized by *G. rarus*, leading to higher acute toxicity over a short duration. In contrast, the slower metabolism and clearance of nonylphenol in *G. rarus* contributes to its higher chronic toxicity under extended exposure. These potential toxicity mechanisms warrant further investigation.

Further collected exposure data from the Web of Science and China National Knowledge Infrastructure (CNKI) for these pollutants found limited results; exposure data in *G. rarus*' habitat (primarily the Tuojiang River basin in Sichuan Province) were only available for 12 PFASs (Supplementary Table S7). These data used liquid chromatography-tandem mass spectrometry (LC-MS/MS) for quantification and reported detection frequencies and concentration ranges. While these data provide a representative snapshot of recent PFAS contamination in the habitat, they may not capture seasonal variations or long-term trends. The presence of these pollutants may be closely linked to industrial activities surrounding *G. rarus*' habitat. Sichuan Province is a key industrial hub in western China and hosts sectors like electronics, automotive, and chemical manufacturing, where PFAS use or emissions are prevalent. There is also a major fluorochemical base near Zigong City. Further, the unique geographical and meteorological conditions of the Sichuan Basin, particularly during winter, may prolong the accumulation of these pollutants in the environment, leading to widespread concern about their environmental impact. The risk assessment of pollutants in *G. rarus*' habitat was conducted based on the toxicity prediction results. The RQ values of acute and chronic toxicity for all identified pollutants to different life stages of *G. rarus* were far below 1, indicating that these pollutants pose a low ecological risk to *G. rarus* (Table 1).

However, it is important to consider their potential long-term ecological risks[51]. Recent studies on PFASs, such as perfluorooctanoic acid (PFOA), have found that these substances exhibit strong persistence in the environment and may bioaccumulate through food chains. This may affect the survival of high trophic-level aquatic organisms, including *G. rarus*[69]. Therefore, it is recommended to conduct long-term dynamic monitoring of the contamination status of PFASs in *G. rarus*' habitat, particularly the spatial distribution and bioaccumulation patterns of substances like PFOA and its new alternatives.

It should be noted that the RQ values presented are point estimates based on ML-QSAR predictions and do not account for prediction uncertainty. However, even considering potential prediction errors (RMSEex = 0.51 for acute and 2.70 for chronic models), all RQ values remain well below 1, supporting the robustness of our low-risk conclusion. Future monitoring efforts should focus on compounds with RQ values approaching 0.01 to reduce uncertainty in risk characterization.

Table 1 Acute and chronic ecological risk entropy associated with 12 pollutants impacting *G. rarus* in different life stages in its habitat

Chemical	Acute risk (RQ)				Chronic risk (RQ)			
	Life stage_1	Life stage_2	Life stage_3	Average	Life stage_1	Life stage_2	Life stage_3	Average
PFBA	2.64E-07	2.65E-07	2.41E-07	2.57E-07	1.98E-06	1.98E-06	1.98E-06	1.98E-06
PFPeA	2.23E-09	2.23E-09	1.99E-09	2.15E-09	1.59E-08	1.59E-08	1.59E-08	1.59E-08
PFHxA	4.26E-09	4.28E-09	3.55E-09	4.00E-09	3.17E-08	3.17E-08	3.17E-08	3.17E-08
PFHpA	6.60E-09	6.63E-09	5.35E-09	6.13E-09	3.74E-08	3.73E-08	3.73E-08	3.74E-08
PFOA	9.40E-05	9.43E-05	6.03E-05	7.93E-05	3.93E-04	3.93E-04	3.93E-04	3.93E-04
PFNA	1.24E-09	1.25E-09	8.04E-10	1.05E-09	3.84E-09	3.84E-09	3.84E-09	3.84E-09
PFDA	9.94E-10	9.97E-10	6.43E-10	8.42E-10	3.06E-09	3.06E-09	3.06E-09	3.06E-09
PFDoDA	3.23E-09	3.23E-09	2.09E-09	2.73E-09	9.95E-09	9.95E-09	9.95E-09	9.95E-09
PFBS	9.93E-10	9.95E-10	6.43E-10	8.41E-10	3.07E-09	3.07E-09	3.07E-09	3.07E-09
PFHxS	2.23E-10	2.19E-10	1.60E-10	1.96E-10	7.19E-09	7.19E-09	7.19E-09	7.19E-09
PFOS	6.23E-08	6.25E-08	4.02E-08	5.27E-08	2.58E-07	2.58E-07	2.58E-07	2.58E-07
PFDS	6.78E-10	6.78E-10	4.99E-10	6.06E-10	1.91E-08	1.90E-08	1.90E-08	1.90E-08

Limitations and future directions

This study has several limitations. First, the models were developed primarily for organic compounds and organometallics; predictions for ionic metals rely on missing value handling by Random Forest and may carry higher uncertainty. Second, the chronic toxicity dataset has an imbalanced life stage distribution (50.7% for adult data), which may reduce sensitivity to life stage differences in chronic exposure. Third, the molecular descriptors used may not capture all relevant toxicity mechanisms, particularly for new contaminants with novel structures. Future research should focus on expanding the dataset as new toxicity data becomes available, incorporating additional descriptors tailored to new contaminants, conducting long-term monitoring across seasons to better characterize exposure variability, and investigating mixture toxicity effects to move toward more realistic risk assessment.

Conclusions

This study integrated multiple molecular descriptors and life stage information to develop an ML-QSAR predictive model to assess the biological toxicity of *G. rarus*. The outcome provides an alternative and relatively accurate tool for predicting toxicity in rare and endangered species. The feature importance analysis revealed that various molecular descriptors (i.e., atomic properties, functional group distribution, spatial position information, and molecular volume) played significant roles in modeling. Life stage was an important feature in constructing the acute toxicity prediction model. This highlights the need to consider the sensitivity and metabolic capacity of *G. rarus* at different life stages to more accurately assess potential risks from pollutants. The constructed model predicted that nonylphenol and certain PFASs (such as PFOS and PFOA) in the habitat would show significantly higher toxicity than the average of other chemicals. However, further ecological risk assessment found the risk entropy values of PFASs approached safety thresholds, but their potential long-term ecological risks require attention. This study's ML-QSAR predictive model provides a novel approach for predicting toxicity in rare and endangered species for which data are scarce, including *G. rarus*. This provides data support for the ecological risk assessment of various pollutants.

Supporting information

It accompanies this paper at: https://doi.org/10.48130/newcontam-0026-0010.

Author contributions

The authors confirm their contributions as follows: Ying Wang: conceptualization, methodology, funding acquisition, supervision, writing – review & editing; Xin Wang: data collection and analysis, visualization, formal analysis, writing – original draft; Yunchi Zhou: software, validation; Yinghao Cheng: software, validation; Xiaomin Li: writing – review & editing; Xiaolei Wang: writing – review & editing, Yuefei Ruan: writing – review & editing, Zhaomin Dong: writing – review & editing; Wenhong Fan: conceptualization, supervision, writing – review & editing. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Funding

This work was supported by the National Key R&D Program of China (Grant No. 2022YFC3204800), the National Natural Science Foundation of China (Grant No. 42177240), and the Beijing Natural Science Foundation (Grant No. 8242033).

Declarations

Competing interests

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

Author details

[1]School of Materials Science and Engineering, Beihang University, Beijing 100191, China; [2]Publishing House of Electronics Industry, Beijing 100036, China; [3]Nuclear and Radiation Safety Center, Ministry of Ecology and Environment, Beijing 100082, China; [4]State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China; [5]State Key Laboratory of Marine Pollution and Department of Chemistry, City University of Hong Kong, Hong Kong 999077, China; [6]Research Centre for the Oceans and Human Health, City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China; [7]Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing 21009, China

References

[1] WWF. 2024. *Living Planet Report 2024 – a system in Peril*. WWF, Gland, Switzerland. https://wwflpr.awsassets.panda.org/downloads/2024-living-planet-report-a-system-in-peril.pdf

[2] Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, et al. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. *Science* 344:1246752

[3] Sigmund G, Ågerstrand M, Brodin T, Diamond ML, Erdelen WR, et al. 2022. Broaden chemicals scope in biodiversity targets. *Science* 376:1280

[4] Qian MM, Wang ZY, Zhou Q, Wang J, Shao Y, et al. 2023. Environmental DNA unveiling the fish community structure and diversity features in the Yangtze River basin. *Environmental Research* 239:117198

[5] Mei Z, Cheng P, Wang K, Wei Q, Barlow J, et al. 2020. A first step for the Yangtze. *Science* 367:1314

[6] Yang H, Shen L, He Y, Tian H, Gao L, et al. 2024. Status of aquatic organisms resources and their environments in Yangtze River system (2017–2021). *Aquaculture and Fisheries* 9:833–850

[7] Zhang K, Yang X, Xu M, Lin Q, Kattel G, et al. 2018. Confronting challenges of managing degraded lake ecosystems in the Anthropocene, exemplified from the Yangtze River Basin in China. *Anthropocene* 24:30–39

[8] Jin X, Wang Y, Jin W, Rao K, Giesy JP, et al. 2014. Ecological risk of nonylphenol in China surface waters based on reproductive fitness. *Environmental Science & Technology* 48:1256–1262

[9] Cheng J, Zou H, Li M, Wang J, Wang G, et al. 2023. Morphological and molecular identification of *Dactylogyrus gobiocypris* (Monogenea: Dactylogyridae) on gills of a model fish, *Gobiocypris rarus* (Cypriniformes: Gobionidae). *Pathogens* 12:206

[10] He Y, Wang J, Blanchet S, Lek S. 2012. Genetic structure of an endangered endemic fish (*Gobiocypris rarus*) in the Upper Yangtze River. *Biochemical Systematics and Ecology* 43:214–225

[11] Bai Y, Lian D, Su T, Wang YYL, Zhang D, et al. 2021. Species and life-stage sensitivity of Chinese rare minnow (Gobiocypris rarus) to

chemical exposure: a critical review. *Environmental Toxicology and Chemistry* 40:2680−2692

[12] Liu C, Huang K, Zhang Z, Yang C, Zhang Y, et al. 2022. Growth retardation in silver carp (*Hypophthalmichthys molitrix*) in the middle reaches of the Yangtze River and its possible causal agents. *ACS ES&T Water* 2:2422−2430

[13] Qin Y, Wei Q, Ji Q, Li K, Liang R, et al. 2023. Determining the position of a fish passage facility entrance based on endemic fish swimming abilities and flow field. *Environmental Science and Pollution Research* 30:6104−6116

[14] Singh KP, Gupta S, Kumar A, Mohan D. 2014. Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology. *Chemical Research in Toxicology* 27:741−753

[15] Wu F, Mu Y, Chang H, Zhao X, Giesy JP, et al. 2013. Predicting water quality criteria for protecting aquatic life from physicochemical properties of metals or metalloids. *Environmental Science & Technology* 47:446−453

[16] Dudek AZ, Arodz T, Gálvez J. 2006. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial Chemistry & High Throughput Screening* 9:213−228

[17] La Farré M, Pérez S, Kantiani L, Barceló D. 2008. Fate and toxicity of emerging pollutants, their metabolites and transformation products in the aquatic environment. *TrAC Trends in Analytical Chemistry* 27:991−1007

[18] Zhao Y, Xia Y, Yu Y, Liang G. 2023. QSAR in natural non-peptidic food-related compounds: current status and future perspective. *Trends in Food Science & Technology* 140:104165

[19] Niazi SK, Mariam Z. 2023. Recent advances in machine-learning-based chemoinformatics: a comprehensive review. *International Journal of Molecular Sciences* 24:11488

[20] Staszak M, Staszak K, Wieszczycka K, Bajek A, Roszkowski K, et al. 2022. Machine learning in drug design: use of artificial intelligence to explore the chemical structure–biological activity relationship. *WIREs Computational Molecular Science* 12:e1568

[21] Schmidt S, Schindler M, Faber D, Hager J. 2021. Fish early life stage toxicity prediction from acute daphnid toxicity and quantum chemistry. *SAR and QSAR in Environmental Research* 32:151−174

[22] Cheng Y, Wang Y, Xu Z, Feng C, Dong Z, et al. 2025. Predicting the site-specific toxicity of metals to fishes using a new machine learning-based approach. *Environmental Science & Technology* 59:14881−14891

[23] Cronin MTD. 2006. The role of hydrophobicity in toxicity prediction. *Current Computer - Aided Drug Design* 2:405−413

[24] Vardy DW, Oellers J, Doering JA, Hollert H, Giesy JP, et al. 2013. Sensitivity of early life stages of white sturgeon, rainbow trout, and fathead minnow to copper. *Ecotoxicology* 22:139−147

[25] Yang G, Lv L, Di S, Li X, Weng H, et al. 2021. Combined toxic impacts of thiamethoxam and four pesticides on the rare minnow (*Gobiocypris rarus*). *Environmental Science and Pollution Research* 28:5407−5416

[26] He J, Peng T, Yang X, Liu H. 2018. Development of QSAR models for predicting the binding affinity of endocrine disrupting chemicals to eight fish estrogen receptor. *Ecotoxicology and Environmental Safety* 148:211−219

[27] O'Boyle NM. 2012. Towards a Universal SMILES representation - a standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics* 4:22

[28] Karade D. 2021. Custom ML module of AIDrugApp for molecular identification, descriptor calculation, and building ML/DL QSAR models. *ChemRxiv* preprint

[29] Consonni V, Todeschini R. 2009. Molecular descriptors. In *Recent Advances in QSAR Studies*, eds Puzyn T, Leszczynski J, Cronin M. Vol 8. Dordrecht: Springer. pp. 29–102 doi: 10.1007/978-1-4020-9783-6_3

[30] Yang Y, Wand J, Sun W. 2023. Screening of antitumor drug based on the combination of featured structure description. *Journal of East China University of Science & Technology* 49(6):907−914 (in Chinese)

[31] Chan JY, Leow SMH, Bea KT, Cheng WK, Phoong SW, et al. 2022. Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics* 10:1283

[32] Chapman SJ. 2018. Review of discovering statistics using IBM SPSS statistics, 4th edition. *Journal of Political Science Education* 14:145−147

[33] Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6

[34] Zhou Y, Wang Y, Peijnenburg W, Vijver MG, Balraadjsing S, et al. 2024. Application of machine learning in nanotoxicology: a critical review and perspective. *Environmental Science & Technology* 58(34):14973−14993

[35] Pes B. 2021. Learning from high-dimensional and class-imbalanced datasets using random forests. *Information* 12:286

[36] Hastie TJ, Pregibon D. 2017. Generalized linear models. In *Statistical Models in S*. London: Routledge. pp. 195–247

[37] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. 2020. A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing* 408:189−215

[38] Zhang S, Li X, Zong M, Zhu X, Wang R. 2018. Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems* 29:1774−1785

[39] Prieto A, Prieto B, Ortigosa EM, Ros E, Pelayo F, et al. 2016. Neural networks: an overview of early research, current frameworks and new challenges. *Neurocomputing* 214:242−268

[40] Song L, Langfelder P, Horvath S. 2013. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics* 14:5

[41] Nti IK, Nyarko-Boateng O, Aning J. 2021. Performance of machine learning algorithms with different K values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science* 13:61−71

[42] Suvannang N, Preeyanon L, Malik AA, Schaduangrat N, Shoombuatong W, et al. 2018. Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study. *RSC Advances* 8:11344−11356

[43] Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, et al. 2012. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 17:4791−4810

[44] Giddings J, Hendley P, Jackson S, Dobbs M, Barefoot A, et al. Aquatic risk assessment of agricultural and residential uses of pyrethroid insecticides in the US: an overview. *Proc. 248th National Meeting. Abstracts of Papers of the American Chemical Society, USA, 2014*. Washington, DC, USA

[45] An JX, Ma Y, Zhao WB, Hu YM, Wang YR, et al. 2023. Drug repurposing strategy II: from approved drugs to agri-fungicide leads. *The Journal of Antibiotics* 76:131−182

[46] Strzelecka M, Świątek P. 2021. 1,2,4-Triazoles as important antibacterial agents. *Pharmaceuticals* 14:224

[47] Díaz-Cruz MS, Barceló D. 2006. Determination of antimicrobial residues and metabolites in the aquatic environment by liquid chromatography tandem mass spectrometry. *Analytical and Bioanalytical Chemistry* 386:973−985

[48] Šudoma M, Neuwirthová N, Hvězdová M, Svobodová M, Bílková Z, et al. 2019. Fate and bioavailability of four conazole fungicides in twelve different arable soils–effects of soil and pesticide properties. *Chemosphere* 230:347−359

[49] Björklund S, Weidemann E, Jansson S. 2023. Emission of per- and polyfluoroalkyl substances from a waste-to-energy plant—occurrence in ashes, treated process water, and first observation in flue gas. *Environmental Science & Technology* 57(27):10089−10095

[50] Joudan S, Lundgren RJ. 2022. Taking the "F" out of forever chemicals. *Science* 377:816−817

[51] Dickman RA, Aga DS. 2022. A review of recent studies on toxicity, sequestration, and degradation of per- and polyfluoroalkyl substances (PFAS). *Journal of Hazardous Materials* 436:129120

[52] Dimond JB, Owen RB. 1996. Long-term residue of DDT compounds in forest soils in Maine. *Environmental Pollution* 92:227−230

[53] Kadiru S, Patil S, D'Souza R. 2022. Effect of pesticide toxicity in aquatic environments: a recent review. *International Journal of Fisheries and Aquatic Studies* 10:113−118

[54] Zhou Y, Wang Y, Peijnenburg W, Vijver MG, Balraadjsing S, et al. 2023. Using machine learning to predict adverse effects of metallic

nanomaterials to various aquatic organisms. *Environmental Science & Technology* 57:17786–17795
[55] Speiser JL, Miller ME, Tooze J, Ip E. 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications* 134:93–101
[56] Khazaee M, Guardian MGE, Aga DS, Ng CA. 2020. Impacts of sex and exposure duration on gene expression in zebrafish following perfluorooctane sulfonate exposure. *Environmental Toxicology and Chemistry* 39:437–449
[57] Moriwaki H, Tian YS, Kawashita N, Takagi T. 2018. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* 10:4
[58] Liu S, Qiu WH, Li RZ, Chen B, Wu X, et al. 2023. Perfluorononanoic acid induces neurotoxicity via synaptogenesis signaling in zebrafish. *Environmental Science & Technology* 57:3783–3793
[59] Sloman KA, McNeil PL. 2012. Using physiology and behaviour to understand the responses of fish early life stages to toxicants. *Journal of Fish Biology* 81:2175–2198
[60] Rønnestad I, Yúfera M, Ueberschär B, Ribeiro L, Sæle Ø, et al. 2013. Feeding behaviour and digestive physiology in larval fish: current knowledge, and gaps and bottlenecks in research. *Reviews in Aquaculture* 5:S59–S98
[61] Nunn AD, Tewson LH, Cowx IG. 2012. The foraging ecology of larval and juvenile fishes. *Reviews in Fish Biology and Fisheries* 22:377–408
[62] Long XB, Yao CR, Li SY, Zhang JG, Lu ZJ, et al. 2024. Screening androgen receptor agonists of fish species using machine learning and molecular model in NORMAN water-relevant list. *Journal of Hazardous Materials* 468:133844
[63] Chen Z, An F, Zhang Y, Liang Z, Liu W, et al. 2023. Single- atom Mo–Co catalyst with low biotoxicity for sustainable degradation of high-ionization-potential organic pollutants. *Proceedings of the National Academy of Sciences of the United States of America* 120:e2305933120
[64] Barouki R. 2010. Linking long-term toxicity of xeno-chemicals with short-term biological adaptation. *Biochimie* 92:1222–1226
[65] Meador JP, Escher B. 2025. Fish early-life stage toxicity and environmental relevance: what does high-dose toxicity testing tell us? *Environmental Toxicology and Chemistry* 44(5):1222–1227
[66] Qin W, Henneberger L, Glüge J, König M, Escher BI. 2024. Baseline toxicity model to identify the specific and nonspecific effects of per- and polyfluoroalkyl substances in cell-based bioassays. *Environmental Science & Technology* 58:5727–5738
[67] Liu S, Zhou J, Guo J, Xue M, Shen L, et al. 2023. Impact mechanisms of humic acid on the transmembrane transport of per-and polyfluoroalkyl substances in wheat at the subcellular level: the important role of slow-type anion channels. *Environmental Science & Technology* 57:8739–8749
[68] Di QN, Cao WX, Xu R, Lu L, Xu Q, et al. 2018. Chronic low-dose exposure of nonylphenol alters energy homeostasis in the reproductive system of female rats. *Toxicology and Applied Pharmacology* 348:67–75
[69] Li Y, Yao J, Pan Y, Dai J, Tang J. 2023. Trophic behaviors of PFOA and its alternatives perfluoroalkyl ether carboxylic acids (PFECAs) in a coastal food web. *Journal of Hazardous Materials* 452:131353