Science of the Total Environment 887 (2023) 164072

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

In silico prediction of acute chemical toxicity of biocides in marine crustaceans using machine learning

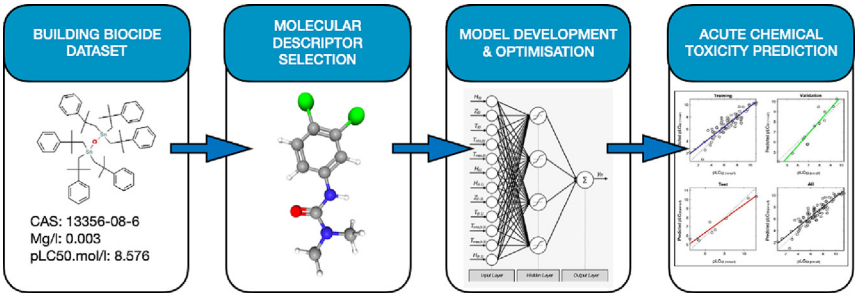
Rama Krishnan [a], Ian S. Howard [b], Sean Comber [c], Awadhesh N. Jha [a,*]

[a] School of Biological and Marine Sciences, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK
[b] School of Engineering, Computing and Mathematics, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK
[c] School of Geography, Earth and Environmental Sciences, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK

HIGHLIGHTS

- Machine Learning models applied for the first time to classify biocide toxicities
- Evaluation of six models to predict toxicities in marine crustaceans
- All the models used showed good predictive performance
- Artificial neural network and decision tree showed the best predictive performance
- ALOGP, SRW10 and SMR molecular descriptors most important to predict acute toxicity

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Henner Hollert

Keywords:
Biocides
LC_{50}
Machine learning
QSAR
Marine crustaceans
Ecotoxicology

ABSTRACT

Biocides are a heterogeneous group of chemical substances intended to control the growth or kill undesired organisms. Due to their extensive use, they enter marine ecosystems *via* non-point sources and may pose a threat to ecologically important non-target organisms. Consequently, industries and regulatory agencies have recognized the ecotoxicological hazard potential of biocides. However, the prediction of biocide chemical toxicity on marine crustaceans has not been previously evaluated. This study aims to provide *in silico* models capable of classifying structurally diverse biocidal chemicals into different toxicity categories and predict acute chemical toxicity (LC_{50}) in marine crustaceans using a set of calculated 2D molecular descriptors. The models were built following the guidelines recommended by the OECD (Organization for Economic Cooperation and Development) and validated through stringent processes (internal and external validation). Six machine learning (ML) models were built and compared (linear regression: LR; support vector machine: SVM; random forest: RF; feed-forward backpropagation-based artificial neural network: ANN; decision trees: DT and naïve Bayes: NB) for regression and classification analysis to predict toxicities. All the models displayed encouraging results with high generalisability: the feed-forward-based backpropagation method showed the best results with determination coefficient R^2 values of 0.82 and 0.94, respectively, for training set (TS) and validation set (VS). For classification-based modelling, the DT model performed the best with an accuracy (*ACC*) of 100 % and an area under curve (*AUC*) value of 1 for both TS and VS. These models showed the potential to replace animal testing for the chemical hazard assessment of untested biocides if they fall within the applicability domain of the proposed models. In general, the models are highly interpretable and robust, with good predictive performance.

Abbreviations: ACC, Accuracy; AD, Applicability Domain; ANN, Artificial Neural Network; AUC, Area Under Curve; BSS, Best Subset Selection; CAS, Chemical Abstracts Service; COMBASE, COMputational tool for the assessment and substitution of Biocidal Active Substances of Ecotoxicological Concern; DT, Decision Trees; ECHA, European Chemicals Agency; ERA, Environmental Risk Assessment; LC50, Lethal Concentration/ Acute Chemical Toxicity; MAE, Mean Absolute Error; MCC, Matthews Correlation Coefficient; ML, Machine Learning; NB, Naïve Bayes; OECD, Organization for Economic Cooperation and Development; PCA, Principal Component Analysis; QSAR, Quantitative Structure-Activity Relationship; RF, Random Forest; RMSD, Root Mean Square Deviation; RMSE, Root Mean Square Error; ROC, Receiver Operating Characteristic; SMILES, Simplified Molecular Input Line Entry System; SVM, Support Vector Machine; TS, Training Set.

* Corresponding author.
E-mail address: a.jha@plymouth.ac.uk (A.N. Jha).

http://dx.doi.org/10.1016/j.scitotenv.2023.164072
Received 21 January 2023; Received in revised form 24 April 2023; Accepted 7 May 2023
Available online 31 May 2023

The models also displayed a trend indicating that toxicity is largely influenced by factors such as lipophilicity, branching, non-polar bonding and saturation of molecules.

1. Introduction

Biocides are a heterogeneous group of chemicals which are used "with the intention of destroying, deterring, rendering harmless, preventing the action of, or otherwise exerting a controlling effect on, any harmful organism by any means other than mere physical or mechanical action" (EU, 2012). These biocides are comprised of an "active substance" incorporated with "co-formulants" (such as stabilizers, solvents, carriers and wetting agents) to ensure the final potency of biocidal mixture (Marzo et al., 2020). These biocidal products, via point and non-point sources, enter the aquatic environments and may pose a threat to ecologically and commercially important non-target organisms with long-term impact on the ecosystems, and human health (Coors et al., 2018; Flemming et al., 2009). For example, in Europe, biocidal products are regulated by the BPR, Regulation (EU: 528/2012) (EU, 2012). According to the current biocidal product regulation (EU, 2012), the formulation, including both "active substance" and "co-formulants", must undergo an environmental risk assessment (ERA) to evaluate the toxicity of biocidal products (Backhaus et al., 2013). Moreover, this regulation improves the efficiency of internal market harmonizing rules and ensures effective protection of the animals and human health and the environment. Additionally, the European Chemicals Agency (ECHA) also ensures the overall applicability and robustness of the legislation by providing technical and scientific support to the European Commission (EC) (EC, 2018). The biocides can be classified into 22 product types (PT) (Marzo et al., 2020) which are further categorized into four groups (Khan et al., 2019). The active substance specific to the PTs also determines their approval.

There are official risk assessment reports by the EC addressing various ecotoxicological risks caused by the use of specific PTs (EC, 2009). The reports suggest that the biocides can be carried away to non-target sites during their applications (e.g., during rain *via* runoff), including the surface water, signifying a threat to the aquatic ecosystem. Sustainable use of biocides is therefore imperative. It is also necessary to emphasize the need to understand the short and long-term consequences of biocides on the aquatic ecosystem and the valuable resources therein. Consequently, in 2016, the EC initiated the LIFE-COMBASE project (COMBASE, 2016). The project aims to promote and encourage the sustainable use of biocides by analyzing the overall risks they pose to the environment and human health. The LIFE-COMBASE project also promotes chemical hazard assessment using alternative methods to animal testing by incorporating *in silico* approaches. The introduction of an innovative approach for environmental health monitoring using the application of machine learning (ML) has recently attracted attention in ecotoxicological studies. The implementation of ML in this context is based on the use of algorithms allowing the system to learn, interpret, and predict the chemical and biological processes associated with it (Miller et al., 2018). With the advancement in these computational approaches, such as read-across (RA) and quantitative structure-activity relationships (QSARs), ML facilitates efficient risk management by eliminating and outperforming unnecessary testing on animals while less time-consuming concurrently (Liu et al., 2018; Miller et al., 2018). A plethora of studies are available reporting that ML approaches in QSAR surpass other computation-based conventional approaches, for instance, knowledge-based functions of datasets and empirical scoring methodologies (Sieg et al., 2019; Barros et al., 2020). Nevertheless, understanding the underlying science and rationale behind selecting features, algorithms and interpretation knowledge is crucial (Sieg et al., 2019; Barros et al., 2020).

Reports suggest that the saltwater habitat is the ultimate sink of numerous biocides and anthropogenic pollutants (Dale and Beyeler, 2001; Liu et al., 2019; Oberdörster and Cheek, 2001). However, to the extent of our knowledge, no published studies are available reporting predictive ML models for environmentally sensitive marine invertebrates such as marine crustaceans for the toxicological evaluation of biocides. Crustaceans such as mysids have been used as model species for nearly two decades as an important tool for toxicity regulation. Mysids represent shrimp-like small crustaceans found in both saltwater and freshwater environments, are an ecologically important group of organisms. In this context, for example, *Americamysis bahia* has served as an ideal species for estuarine and coastal monitoring by the American Society for Testing of Materials (ASTM) and US-Environmental Protection Agency (US-EPA) (Langdon et al., 1996; Lussier et al., 1999; Roast et al., 1999).

In the light of above information, our study aimed to build highly predictive and robust *in silico* models. These models were validated through stringent processes to probe the acute chemical toxicity of various biocides on marine crustaceans. In order to achieve the objectives, firstly, an acute chemical toxicity or LC_{50} dataset was built, which is the mean lethal concentration, determining the concentration of a substance in the medium causing mortality to 50 % of a group of test organisms within a period of exposure (Rand, 1985). The toxicity data were generated for the three families of marine crustaceans, including *Mysidae*, *Palaemonidae* and *Penaeidae*. Subsequently, regression and classification-based computational models were built to predict the biocide toxicity in these marine crustaceans. In these predictive models, the chemicals were represented as molecular descriptors. Following this, the key molecular descriptors influencing acute chemical toxicity were investigated using ML methods. The molecular descriptors were also employed to check the applicability domain of the chemicals in the dataset.

2. Materials and methods

2.1. Dataset sources

In order to build the biocide acute chemical toxicity (i.e., LC_{50}) dataset for marine crustaceans, firstly, a list of biocides was retrieved from the ECHA (2022), published on 14 May 2022. Secondly, a chemically heterogenous LC_{50} value dataset ($n = 2165$) towards the three families of marine crustaceans (viz. *Mysidae, Palaemonidae* and *Penaeidae*) were downloaded using the US-EPA ECOTOX database (Olker et al., 2022), and the values with an experimental observation time of four days (published 16 May 2022) was selected. Thirdly, the biocidal compounds from the LC_{50} dataset were manually selected. The biocide identification (i.e., Chemical Abstracts Service; CAS and chemical names) was manually compared and retrieved from PubChem (Kim et al., 2021) to circumvent any error in the dataset. Subsequently, the SMILES (simplified molecular input line entry system) strings were converted from chemical structures of biocides for further molecular representation using python script and ChemSpider website (https://www.chemspider.com).

2.2. Dataset pre-processing

For modelling purposes and to improve the overall performance of ML models, the compounds with incorrect CAS numbers or molecular structures not clearly identified were removed from the dataset. Furthermore, to retain an uniformity of biocides in the dataset, metal complexes, inorganic compounds, mixtures with unknown compositions, and salts containing organic counterions were removed. Additionally, the structure of the remaining salts in the dataset was also neutralized. From the dataset containing biocides to be used for modelling, all LC_{50} units were first converted to parts per million (ppm) and data with units that could not be directly converted, for example, AI (active ingredient) ppm, AI μg/l, and mol/l, were removed. Later, any duplicates were removed, and the geometric mean of similar compounds with multiple experimental values were

calculated. Finally, the observed values expressed as ppm (or mg/l) were converted to mmol/l followed by negative logarithmic transformation ($-\text{Log } 10_{\text{ mmol/l}}$) or p-transformation, i.e., pLC_{50}, in accordance with ecotoxicological QSAR studies. The purpose of p-transformation is to reduce the skewness of the data, which can be beneficial for statistical analysis that assume normally distributed data. Consequently, higher pLC_{50} values corresponded to higher toxicity and *vice versa*.

For classification modelling, the guidelines provided by the US-EPA were followed, which suggests classifying the different toxicity categories of chemicals for ecological risk assessment. Accordingly, the chemical aquatic toxicity (ppm) can be classified into five categories, i.e., very highly toxic (<0.1), highly toxic (0.1–1), moderately toxic (>1–10), slightly toxic (>10–100), and non-toxic (>100) (US-EPA, 2021).

2.3. Calculation of molecular descriptor

Molecular descriptors are defined as the numeric representation of various molecular properties derived using mathematical algorithms (Mauri and Srl, 2021). These mathematical representations of molecular descriptors are used to quantitatively represent several chemical and physical characteristics of the molecules. For instance, the lipophilicity of a molecule is quantitatively represented as the molecular descriptor LogP (Chandrasekaran et al., 2018). The molecular descriptors can be categorized into multiple groups based on the dimensionality of the molecular structure, such as 0- to 3-dimensional (3D) descriptors (Mauri and Srl, 2021).

To avoid any conformational complexity and for ease of interpretability, only 2D molecular descriptors were calculated in this study. These molecular descriptors were retrieved from the 2D characterization of molecular structures, which quantify the molecular characteristics such as connectivity of atoms in a molecule and atomic composition (Mauri and Srl, 2021). Firstly, the SMILE strings for each molecule were created, which are the linear structural concepts describing the structure of chemical species. Secondly, in total, 2223 molecular descriptors were calculated, comprising of 2D atom pairs, atom type E-state indices, functional group counts, constitutional indices, topological indices, ring descriptors, atom-centred fragment molecular property, and 2D molecular descriptors were calculated using PaDEL2 and Dragon v. 7 from the open access OCHEM database (Sushko et al., 2011). Additionally, the RDKIT 2D molecular descriptors were also calculated using KNIME Analytics Platform version 4.3.1 (Berthold et al., 2009).

2.4. Feature selection and dataset division

In order to improve the overall generalisability and predictive performance, various feature selection methods were employed, which utilised the most appropriate and relevant features (molecular descriptors) to train the model by eliminating noise in the data. From the initial pool of 2223 features calculated for each chemical, first, the dataset was divided randomly into a training set and test set (80:20 ratio) using an R-script, and only the training set was subjected to feature selection to avoid any bias during model selection. Subsequently, above 80 % zero values and inter-correlated features (>0.90) were eliminated from the dataset using *nearZeroVar* and *findCorrelation* function in RStudio (Kuhn, 2008). Secondly, for regression analysis, the XGBoost modelling approach was applied and validated using 10-fold cross-validation in python3 to select the twenty features with the highest importance (Chen and Guestrin, 2016). Finally, out of the twenty selected features, the Best Subset Selection (BSS) method was employed in python3, which determined the best subset of ten features that best described the endpoints.

2.5. Diversity in the dataset

To develop a robust model with high accuracy and reliable predictions, it is crucial that the chemicals in the dataset are diverse. The diversity of chemicals in our dataset was investigated by first calculating Morgan (2D circular) fingerprints of radius 2 and 1024 nBits for each chemical. The rationale behind selecting the specific fingerprint can be found in previous studies (Kensert et al., 2018; Liu et al., 2019). Secondly, the Tanimoto similarity index was calculated, which can be explained by the equation: $S_{A,B} = c/[a + b - c]$ and $S = 1/(1 + \text{distance})$, where S denotes similarities, a and b represent the number of bits in molecule A and B, respectively; while c represents the number of bits that are in both molecules. Lastly, a heatmap was created to compare the similarities of each chemical. The entire process was performed using KNIME v 4.3.1 (Berthold et al., 2009). In addition, principal component analysis (PCA) was also implemented to define the chemical space occupied by the compounds and diversity in the dataset. The PCA analysis takes the high-dimensional sets of correlated molecular properties or molecular descriptors into consideration and combines them to create a lower-dimensional space of the corresponding properties making it easier to illustrate and interpret the molecular diversity (Walters, 2019).

2.6. Model building

For regression models, four supervised ML algorithms were employed, which are random forest (RF), artificial neural network (ANN), linear regression (LR), and support vector machine (SVM). In supervised learning, the algorithm is trained using "labelled" datasets and the prediction/classification is based on the data provided (Yao et al., 2018).

The SVM, LR and RF algorithms were implemented in Orange v 3.26.0 (Demšar et al., 2013), and the dataset was split into subsets so that 62 compounds (80 %) were used to train the model (training set) and 17 compounds (20 %) were used to test the model (test set). In the case of ANN, feed-forward backpropagation method was employed using Neural Net Fitting app in MATLAB R2021a (MATLAB, 2010) and the model was trained using the Levenberg-Marquardt technique. The dataset was split into 67 compounds (75 %) as a training set, 13 compounds (15 %) as validation set

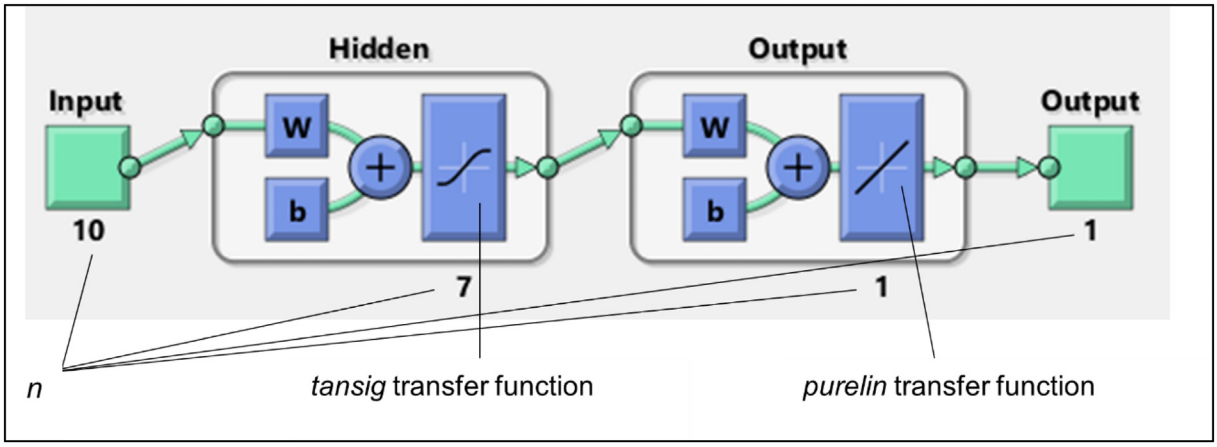


Fig. 1. ANN architecture used for model building (n = no. of neurons used in each layer, w = weight vector and b = bias).

Table 1
Machine Learning (ML) modelling approaches used in this study.

Analysis	Model	Equation	Hyperparameter	Reference
Regression	SVM	$K(X_1, X_2) = \exp\left(-\frac{\|X_1 - X_2\|^2}{2\sigma^2}\right)$	• RBF Kernel	Chang et al., 2010
	RF	$\widehat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x^{\|})$	• No. of trees: 10 • No. of attributes in each split: 8	Ho, 1995
	LR	$Y_i = f(X_i, \beta) + e_i$	• Lasso regression • $\alpha = 0.0001$	Cohen et al., 2013
	ANN	$g(x) = f^L(W^L f^{L-1}(W^{L-1} \ldots f^1(W^1 x) \ldots))$	• Method: Backpropagation • Training: Levenberg-Marquardt	Tahmasebi & Hezarkhani, 2011
Classification	DT	$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$	• Split criterion: Gini diversity index • Max no. of splits: 4–00	Gini, 1936
	NB	$P(x_y\|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$	• Kernel type: Gaussian	Rennie et al., 2003

RBF – radial basis function, σ- variance, X1 and X2 – two points, K – kernel function, B – bagging, $x^{|}$- test samples, b = 1, f_b- trees, Y_i- dependent variable, f- function, X_i- independent variable, β- unknown parameters, e_i- error terms, x – input, y – output, f^L- ReLU function, L – no. of layers, W^L- the weights between layer l − 1, C – branch, σ- independent variable.

and 9 compounds (10 %) as test set. The ANN model consisted of one input layer with ten neurons (number of features), one hidden layer consisting of seven neurons (iteratively tuned and configured for best performance) and one output layer consisting of one neuron. The Tan-Sigmoid transfer function (*tansig*) was employed in the hidden layer, while for the output layer, the Linear Transfer function (*purelin*) was employed. The architecture used to build the ANN model is illustrated in Fig. 1. Similarly, for classification modelling, two supervised ML algorithms were employed, which are decision tree (DT) and naïve Bayes (NB). These algorithms were implemented in MATLAB R2021a (MATLAB, 2010). The details of these ML algorithms and configurations are mentioned in Table 1. More theoretical and mathematical details can be found in previous studies (Liu et al., 2019; Miller et al., 2019; Russom et al., 1997; Schüürmann et al., 2011; Singh et al., 2013).

2.6.1. Validation and performance evaluation

The k-fold cross-validation method was employed to evaluate the robustness and prediction accuracy of each model used while training for both regression and classification analysis. In addition, a test set for external validation was also provided. The number of k in k-fold cross-validation was determined by comparing the predictive performance and multiple iterations. For instance, in the 10-fold cross-validation process, the training set was randomly divided into ten subsets, out of which nine subsets were randomly used as the training set. The remaining subset was used as the test set to evaluate the predictive accuracy (Arlot and Celisse, 2010). The cross-validation method was repeated 100 times to maximize reliability and minimize the possibilities of error. For ML model analysis, the predictive performance was evaluated by the following statistical estimators: mean absolute error (*MAE*), coefficient of determination (R^2), root-mean-square deviation (*RMSD*) or root-mean-square error (*RMSE*), mean squared error (*MSE*), an area under curve (*AUC*), specificity (*SP*), sensitivity (*SE*), and model accuracy (*ACC*). The details of these statistical algorithms are mentioned in Table 2.

2.6.2. Applicability domain (AD) study

The AD of our ML models was further analyzed to investigate the reliability of the models in accordance with the OECD principle 3 (OECD, 2004). In this study, the standardization approach was employed using the software Applicability Domain v1.0 proposed by Roy et al. (2015) to define our dataset's chemical space and probe outliers present in the training set and test set. The approach firstly follows standardizing descriptors in the developed model (all compounds) using the formulae:

$$S_{ki} = \frac{|X_{k_i} - \overline{X}_i|}{\sigma_{X_i}}$$

where k= total no. of compounds, i= total no. of descriptors, S_{ki}= standardised descriptors, X_{k_i}= original descriptors, $\overline{X}_i$= mean of X_{ki}, σ_{X_i}= standard deviation of X_{k_i} for training set.

Secondly, if $[S_i]_{\max(k)} \leq 3$, then the compound is not an X-outlier or within AD. Else, calculate $[S_i]_{\min(k)} > 3$, which indicates the compound is an X-outlier or outside AD. In the case of $[S_i]_{\max(k)} > 3$ and $[S_i]_{\min(k)} < 3$, $S_{new(k)}$ has to be calculated using the equation:

$$S_{new(k)} = \overline{S}_k + 1.28 \times \sigma_{S_k}$$

where, $S_{new(k)}$ = S_{new} value for compound k, $\overline{S}_k$= mean of $S_{i(k)}$, σ_{S_k} = standard deviation of $S_{i(k)}$.

Hence, if $S_{new(k)} \leq 3$, the compound is not an X-outlier or within AD, and vice versa.

Table 2
Statistical algorithms to estimate the predictive performance of ML models.

Analysis	Statistical estimator	Theory	Equation	Reference
Regression	MSE	Average squared difference between predicted value and actual value	$MSE = \frac{1}{n}\sum_{i=0}^{n}\left(Y_i - \widehat{Y}_i\right)^2$	Bickel & Doksum, 2015
	RMSE/RMSD	Standard deviation of prediction errors	$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \widehat{x_i}\right)^2}{N}}$	Barnston, 1992
	MAE	Deviation of predicted value from the observed value	$MAE = \frac{\sum_{i=1}^{n}\|y_i - x_i\|}{n}$	Willmott & Matsuura, 2005
	R^2	Variation in prediction proposed by the model	$R^2 = 1 - \frac{RSS}{TSS}$	Damodar & Dawn, 2009
Classification	SE	Percentage of positive class predicted as positive	$SE = \frac{TP}{TP+FN}$	Altman & Bland, 1994
	SP	Percentage of negative class predicted as negative	$SP = \frac{TN}{TN+FP}$	Altman & Bland, 1994
	ACC	Fraction of correct prediction to overall predication	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$	Chicco & Jurman, 2020
	AUC	Overall performance of classification model under all classification thresholds	$AUC = \int TPR\, d(FPR)$	Hanley & McNeil, 1982

n - number of data points, Y_i - observed value, $\widehat{Y}$ - predicted value, x_i - observed value, $\widehat{xi}$ - predicted value, N - sample size, y_i - predicted value, x_i - true value, n - total number of data points, RSS – sum of squares of residuals, TSS – total sum of squares, TP – true positive, TN – true negative, FP – false positive, FN – false negative, TPR – true positive rate, FPR – false positive rate.

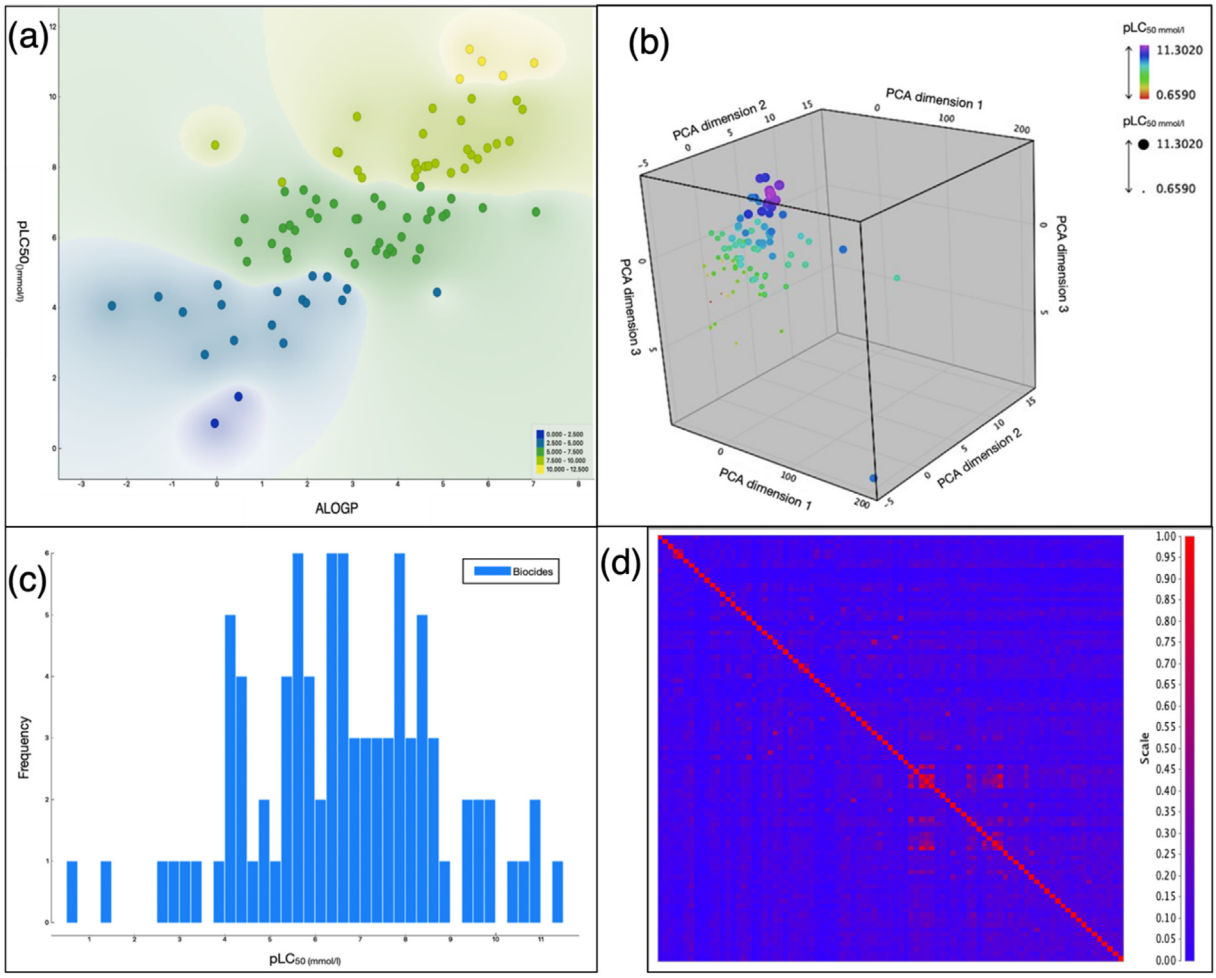


Fig. 2. Figures illustrating diversity in the dataset: (a) ALOGP molecular descriptor correlation with experimental toxicity $pLC_{50\ mmol/l}$. (b) Chemical space of biocide dataset defined using principal component analysis (PCA). The colours and sizes represent the varying $pLC_{50\ mmol/l}$ values of biocides in the dataset. (c) Frequency and distribution of biocides (blue bar) in the marine crustacean toxicity dataset according to their toxicity ($pLC_{50\ mmol/l}$). (d) Tanimoto similarity index heatmap of the biocidal compounds in the dataset using 2D circular Morgan fingerprints. The similarity index increases from zero to one.

3. Results and discussion

3.1. Dataset analysis

The aim of this study was to build QSAR models suitable to predict acute biocide toxicity for marine crustaceans. This was essential since the existing QSAR models provide poor predictive results on marine crustaceans and biocides in particular, as they are trained with diverse chemical datasets. All the biocide LC_{50} datasets for marine crustaceans were collected from the US-EPA ECOTOX database, and the data with an experimental observation time of 96 h or four days were selected. After pruning the dataset with redundant values and standardizing the compounds, the final dataset comprised quite a small set of biocidal compounds ($n = 89$, supplementary file 1). The small number of compounds in the training set and test set limits the overall predictive performance of the models.

The frequency of distribution pattern in our dataset for experimental acute toxicity values ($-\text{Log10}_{\ mmol/l}$), i.e., pLC_{50} of the biocide compounds used for regression and classification modelling was assessed by illustrating a histogram (Fig. 2c). It is to be noted that all the experimental chemical values as ppm or mg/l were converted into mmol/l followed by negative logarithmic transformation ($-\text{Log 10}_{\ mmol/l}$), i.e., pLC_{50} in accord with ecotoxicological QSAR studies. The vertical bars in the histogram represent the occurrence or frequency values of pLC_{50} in the dataset, which were converted into sub-ranges (bins). According to the guidelines by the US-EPA, the dataset was also classified into five categories, i.e., very highly toxic, highly toxic, moderately toxic, slightly toxic, and non-toxic (Table 3). Finally, the dataset was randomly divided in the ratio of 80:20 into a training set and a test set using R script. The training and test sets consisted of 71 and 18 compounds, respectively.

3.2. Diversity analysis in dataset

The diversity of chemical compounds in the dataset was assessed by implementing principal component analysis (PCA) and Tanimoto similarity index. The PCA analysis utilised the molecular descriptors to define a chemical space (Fig. 2b) which is a graphical representation of all the chemicals distributed in a space corresponding to their molecular similarities. Consequently, in this space, the chemicals with similar molecular properties will be close to each other, and chemicals that are distant with their molecular properties will be far apart. Similarly, various dimensions of the PCA analysis (Fig. 2b) showed that the substances in our dataset were clustered, yet good segregation was observed based on the pLC_{50} toxicity values. This is because the dataset comprised the same class of chemicals (biocides) and substances with high pLC_{50} being more prevalent than the rest.

Additionally, the Morgan (2D circular) fingerprints of radius 2 and 1024 nBits were used to construct a Tanimoto similarity heatmap which defined the similarity matrix for each compound (Fig. 2d), where the similarity increased from zero (blue) to one (red). Morgan fingerprints are a type of circular fingerprint that encode molecular structure information as a bit string. They are particularly useful for measuring diversity in a dataset

Table 3
Chemical toxicity categories in marine organisms.

Marine crustacean acute concentration (PPM)	Category used for classification modelling	Binary Classification	Quantity in dataset (n = 89)
<0.1	2	Very highly toxic	64
0.1–1		Highly toxic	
>1–10	1	Moderately toxic	13
>10–100	0	Slightly toxic	12
>100		Nontoxic	

since they capture important structural features of molecules relevant to their biological activity (Rogers and Hahn, 2010). The heatmap revealed that the substance in our dataset was diverse. Overall, the figures (Fig. 2 a-d) illustrate a good diversity of chemicals throughout the dataset.

3.3. Molecular descriptor feature selection and relevance to toxicity prediction

In conjunction with the quality of dataset used, selecting the most relevant molecular descriptors for toxicity prediction is crucial for optimizing the models and unravelling the molecular factors contributing to toxicity. To improve the overall generalisability and to avoid overfitting in our QSAR models, feature selection of the initially calculated molecular descriptors was performed. The features from the initial pool of 2223 molecular descriptors retrieved from Dragon v. 7, PaDEL 2 and RDKiT were reduced using feature selection techniques such as nearZeroVar, findCorrelation, XGBoost and Best Subset Selection (BSS). From the initial pool of 2223 molecular descriptors, 1825 molecular descriptors having >80 % zero values and inter-correlated features (>0.90) were eliminated from the dataset using *nearZeroVar* and *findCorrelation* function in RStudio. From the remaining 398 molecular descriptors, the top 20 were reserved using XGBoost regression modelling in python3, and finally, the top 10 molecular descriptors were selected using the best subset selection (BSS) and used in regression modelling, which are: VE1_Dt, VE2_Dt, B07[C—C], H.049, C.002, ALOGP, XLogP, MLFER_S, SRW10 and SMR. For classification, eighteen descriptors were selected and used by employing XGBoost classification approach in python3 to build the final classification models, which are: Psi_e_1, nRCN, H.049, F01.C.N., F05.N.O., TPSA.NO., ALogP, ATSC1c, ATSC0p, MATS1v, MATS4p, GATS1i, MIC5, JGI6, Chi3v, Chi4v, slogp_VSA10 and smr_VSA3. The XGBoost feature selection for classification modelling works by selecting the most important features and can reduce the noise in the data, making it easier for the algorithm to find meaningful patterns. This often leads to improved model performance, as the algorithm can focus on the most relevant features for the classification task (Devi et al., 2023).

Additionally, to assess the relevancy of the selected molecular descriptors to predict toxicity, the Pearson correlation (*r*) method was employed for the set of molecular descriptors in regression analysis. This method is commonly used to measure the linear relationship between two continuous variables, where the *r*-value ranges from −1 to 1, with −1 indicating a perfectly negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfectly positive linear relationship (Ebenuwa et al., 2019). The *r*-values of the features used for regression were retrieved in the order: ALOGP: +0.703; SRW10: +0.606; SMR: +0.603; VE1_Dt: +0.599; XLogP: +0.578; MLFER_S: +0.410; VE1_Dt: +0.373; H.049: −0.222; C.022: −0.031.

The Pearson correlation statistics suggest that ALOGP describes the pLC_{50} of a chemical best when compared to the rest molecular descriptors. This phenomenon can be justified as ALOGP or Atomic LogP describes the hydrophilicity of a compound. A lower value of LogP suggests higher hydrophilicity of the chemical compound and vice versa. This is because chemicals with high ALOGP value or highly hydrophobic nature tend to remain in the aquatic environment and are ingested and accumulated in the tissues of aquatic organisms (Miller et al., 2019). Furthermore, as illustrated in Fig. 2a, the correlation of ALOGP with toxicity or pLC_{50} suggests that most biocidal substances in our dataset tend to be highly lipophilic.

It is important to note that while the Pearson correlation method is widely used to measure the relevancy of the features, it does have some limitations. Firstly, it only captures linear relationships between variables,

Table 4
Molecular descriptors used for model building.

Model	Descriptors	Software	Description	Descriptor type
Regression	VE1_Dt	Dragon v. 7	Coefficient sum of the last eigenvector from detour matrix	2D matrix-based descriptors
	VE2_Dt		Average coefficient of the last eigenvector from detour matrix	
	B07[C—C]		Presence/absence of C - C at topological distance 7	2D Atom Pairs
	H.049		H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	Atom-centred fragments
	C.002		CH2R2	
	ALOGP		Ghose-Crippen octanol-water partition coeff. (logP)	Molecular Properties
	XLogP	PaDEL 2	octanol/water partition coefficients of organic compounds	XLogP
	MLFER_S		Combined dipolarity/polarizability	Molecular linear free energy relation
	SRW10		Self-returning walk count of order 10 (ln(1 + x))	Walk counts
	SMR	RDKiT	Molecular refractivity	2D
Classification	Psi_e_1	Dragon v. 7	electrotopological state pseudoconnectivity index - type 1	Topological indices
	nRCN		number of nitriles (aliphatic)	Functional group counts
	H.049		H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	Atom-centred fragments
	F01.C.N.		Frequency of C - N at topological distance 1	2D Atom Pairs
	F05.N.O.		Frequency of N - O at topological distance 5	
	TPSA.NO.		topological polar surface area using N,O polar contributions	Molecular Properties
	ALogP	PaDEL 2	Ghose-Crippen LogKow	ALogP
	ATSC1c		Centered Broto-Moreau autocorrelation - lag 1 / weighted by charges	Autocorrelation
	ATSC0p		Centered Broto-Moreau autocorrelation - lag 0 / weighted by polarizabilities	
	MATS1v		Moran autocorrelation - lag 1 / weighted by van der Waals volumes	
	MATS4p		Moran autocorrelation - lag 1 / weighted by van der Waals volumes	
	GATS1i		Geary autocorrelation - lag 1 / weighted by first ionization potential	
	MIC5		Modified information content index (neighbourhood symmetry of 5-order)	Information content
	JGI6		Mean topological charge index of order 6	Topological charge
	Chi3v	RDKiT	Similar to Hall Kier Chi3v, but uses nVal instead of valence	Topochemical descriptors
	Chi4v		Similar to Hall Kier Chi4v, but uses nVal instead of valence	
	slogp_VSA10		MOE logP VSA Descriptor 10 ($0.40 \leq x < 0.50$)	Molecular surface area descriptors
	smr_VSA3		MOE MR VSA Descriptor 3 ($1.82 \leq x < 2.24$)	

meaning it may miss important non-linear relationships. Secondly, it only measures the relationship between two variables at a time, and may not account for the effects of multiple variables on the target variable. To address these limitations, researchers can use more advanced techniques, such as regularisation methods like Lasso or Ridge regression, which can capture non-linear relationships and account for multiple variables simultaneously.

In addition to ALOGP, VE1_Dt and VE2_Dt are molecular descriptors that measure the topological complexity of a molecule. In general, molecules with higher values of VE1_Dt and VE2_Dt tend to be more hydrophobic and less soluble in water, while molecules with lower values tend to be more hydrophilic and more soluble. BO7[C-C] calculates the number of pairs of carbon atoms separated by a distance of 7 or fewer bonds. MLFER_S is a useful molecular descriptor for predicting the solubility of drugs and other bioactive molecules, as solubility is a key factor affecting a drug's bioavailability and pharmacokinetics (Huang et al., 2016). SRW10 is a type of topological descriptor that represents the presence and distribution of various substructures within a molecule. It is useful for QSAR modelling in particular as it captures information about specific substructures that may be important for binding to the target (Hansch and Fujita, 1964).

Other molecular descriptors used to build both regression and classification models have similar properties, while some are different and provide important information about a compound's properties and potential effects on biological systems; their summary has been presented in Table 4. An important point to note here is that the test set was never used during the feature selection process to avoid any kind of bias during model selection.

Table 5
Performance parameters for ANN regression model to predict acute toxicity of biocides.

Model	Dataset	No. of compound	MSE	RMSE	R^2
Feed-forward back propagation	Training set	67	0.89	0.93	0.82
	Validation Set	13	0.46	0.67	0.90
	Test Set	09	0.47	0.68	0.94

3.4. Regression modelling

The regression models to predict the acute toxicity (pLC_{50}) of biocide chemicals were built using our four best-performing modelling approaches (RF, SVM, LR, ANN). The overall generalisability, robustness and predictive performance were determined through stringent internal and external validation procedures. For internal validation, 10-fold cross-validation was employed, whereas, for external validation, a sub-set of the dataset, i.e., a test set (20 %), was used. The criteria to assess the predictive performance and reliability were set using *MSE*, *RMSE*, *MAE* and R^2.

The three-layer feed-forward backpropagation ANN model provided the most satisfactory results compared to other regression models. The model yielded *MSE*, *RMSE* and R^2 values of 0.89, 0.93 and 0.82 in terms of 10-CV; 0.46, 0.67 and 0.90 for the validation set; and 0.47, 0.68 and 0.94 during the external validation using test set (Fig. 3, Table 5). The

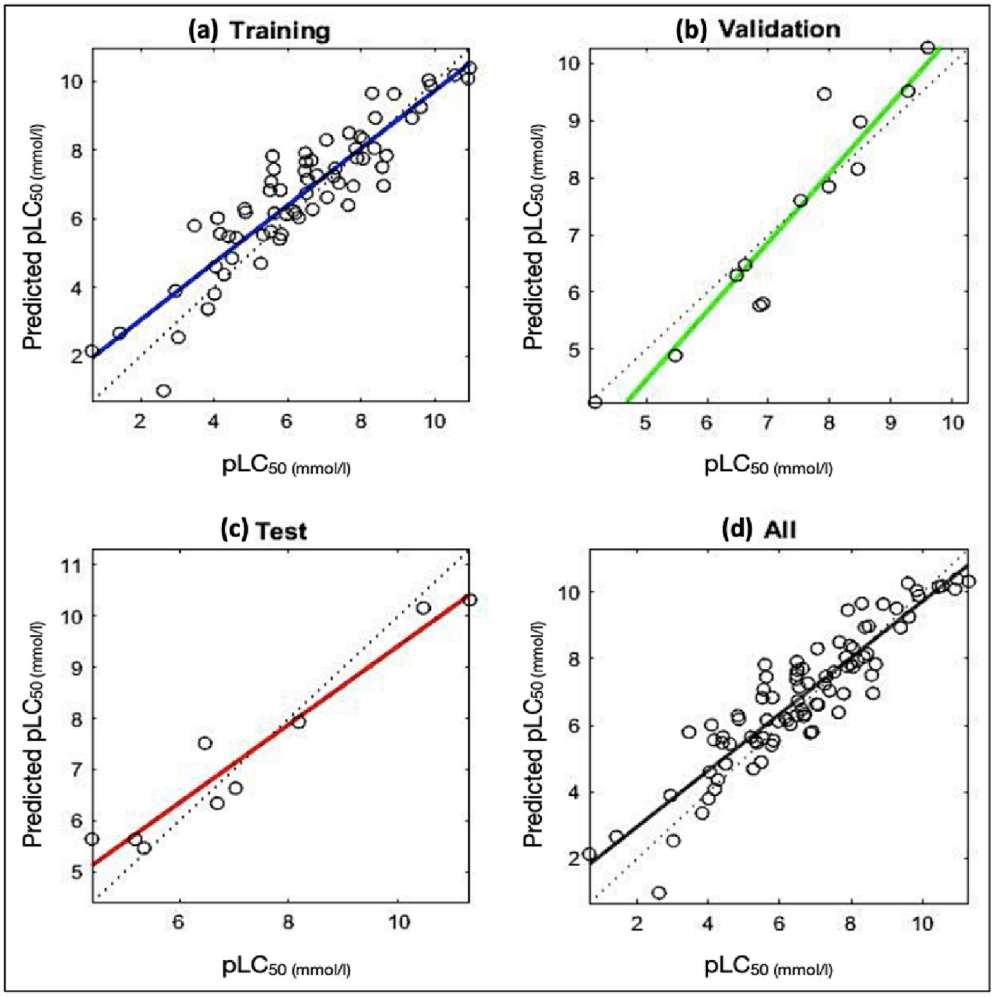


Fig. 3. Scatterplot of the experimented and model predicted values of biocide toxicity (pLC_{50}) in the training set, validation set, test set and complete set of ANN model.

Levenberg-Marquardt (LM) algorithm used to build this model iteratively adjusts the model parameters to minimize the residual sum of squares between the model predictions and the observed data. At each iteration, the algorithm calculates the gradient and Hessian matrix of the objective function (which is the residual sum of squares) and then adjusts the model parameters by solving a modified system of equations that combines the Gauss-Newton method with the steepest descent method (Bilski et al., 2020). This technique results in the overall improvement of the model's generalisability.

In the case of the LR model, the model was obtained in the form of an equation:

$$\begin{aligned} pLC_{50} = {} & 3.25598 - 1.17895\ \text{B07.C.C.} = 0 + 3.97206e - 14\ \text{B07.C.C.} \\ & = 1 - 0.03476\ \text{SMR} - 0.660787\ \text{H.049} + 0.409287\ \text{MLFER\_S} \\ & + 17.1359\ \text{VE1\_Dt} - 262.482\ \text{VE2\_Dt} - 0.0056275\ \text{ALOGP} \\ & + 0.411\,825\ \text{SRW10} - 0.104173\ \text{C.002} + 0.596742\ \text{XLogP} \end{aligned}$$

The LR model yielded satisfactory results for the 10-CV, with *MSE*, *RMSE*, *MAE* and R^2 values of 1.48, 1.22, 0.94 and 0.69, respectively

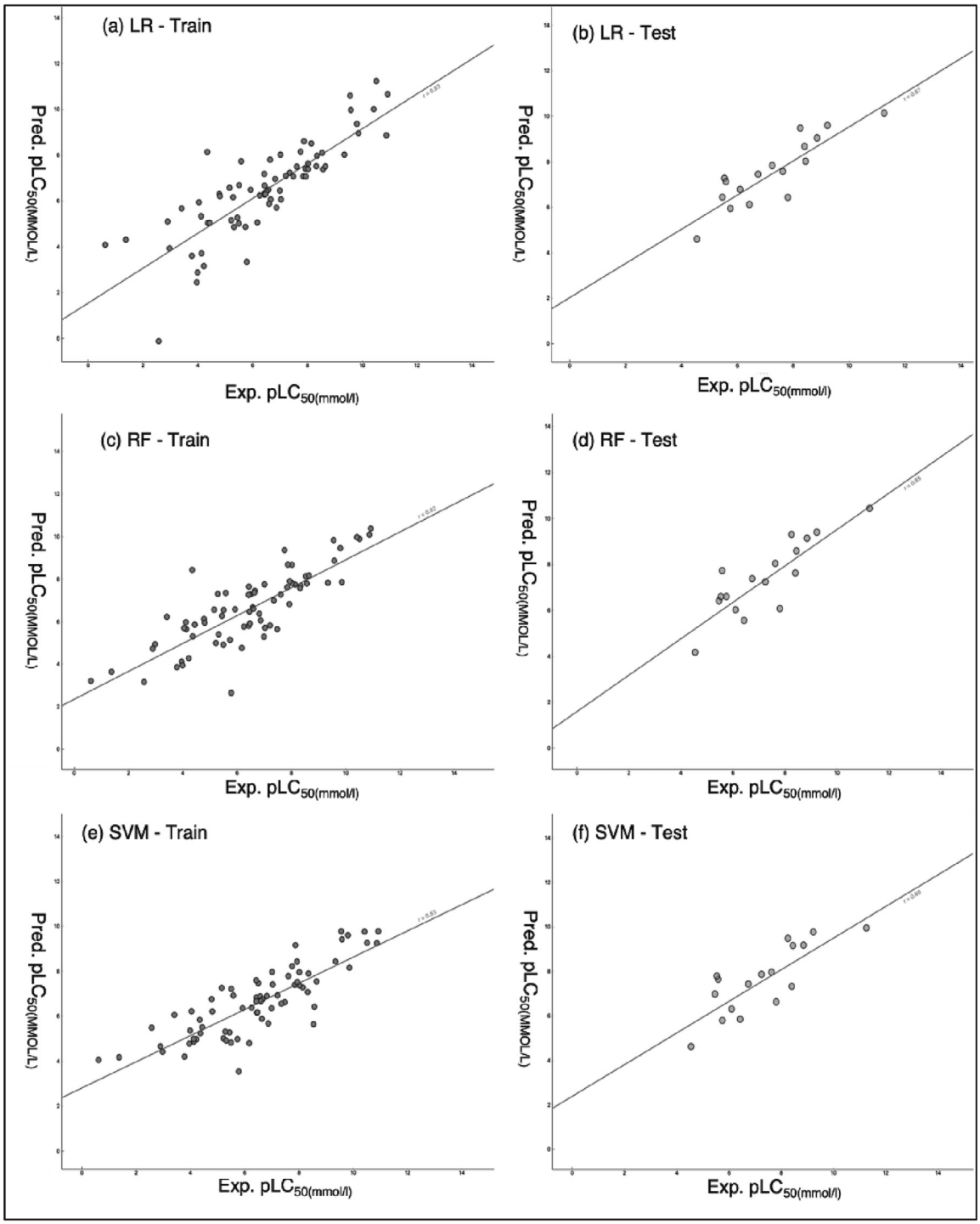


Fig. 4. Regression scatter plots for training and test sets of machine learning models (a–b) LR, (c–d) RF, (e–f) SVM, respectively, used in this study (Experimental pLC_{50} – x-axis vs. Predicted pLC_{50} – y-axis).

Table 6
Performance parameters for various regression models to predict acute toxicity of biocides.

Model	Dataset	MSE	RMSE	MAE	R^2
SVM	Training set	1.56	1.25	0.96	0.69
	Test set	1.08	1.04	0.81	0.64
Random forest	Training set	1.56	1.25	0.97	0.64
	Test set	0.81	0.90	0.70	0.72
Linear regression	Training set	1.48	1.22	0.94	0.69
	Test set	0.70	0.84	0.66	0.76

(Fig. 4a) and performed better during the external validation with *MSE*, *RMSE*, *MAE* and R^2 value of 0.70, 0.84, 0.66 and 0.75, respectively (Fig. 4b). The good predictive performance of the LR model could be due to employing Lasso regression technique, which adds regularisation terms to the cost function to prevent overfitting and improve the generalisability of the model (Yazdi et al., 2021).

In the case of the RF model, the model performed poorly yet satisfactorily compared to LR and ANN models in terms of both 10-CV and external validation. The model yielded the *MSE*, *RMSE*, *MAE* and R^2 values of 1.56, 1.25, 0.97 and 0.67, respectively, for the 10-CV (Fig. 4c) and 0.81, 0.90, 0.70 and 0.71, respectively, during external validation (Fig. 4d). The RF model displayed decent generalisability by constructing ten decision trees and using 8 number of the selected subset of the input data and features. Then the final prediction was made by averaging the predictions of all the individual trees. This approach helps to reduce the risk of overfitting and improves the generalisability of the model (Isabona et al., 2022).

On the other hand, the SVM model displayed slight overfitting on the training set and underperformed compared to the other linear and non-linear regression models yet produced moderate results. The model yielded *MSE*, *RMSE* and R^2 values of 1.56, 1.25, 0.96 and 0.67, respectively, for the 10-CV (Fig. 4e) and 1.08, 1.04, 0.81 and 0.61 during the external validation (Fig. 4f). The possible explanation is, SVM models are particularly susceptible to overfitting when the model has too many features relative to the size of the training data, leading to a sparse and high-dimensional feature space (Han and Jiang, 2014). Another reason could be that model's parameters, such as the regularisation parameter and the kernel function, are not chosen correctly (Han and Jiang, 2014).

Further, the summary and experimented pLC_{50} versus predicted pLC_{50} scatterplots are illustrated in Table 6 and Fig. 4(a–f). An observation made on the measured and predicted biocide toxicity variation pattern in both training and validation sets suggests that all models performed reasonably well.

3.5. Classification modelling

Classification modelling was performed to categorize the biocidal chemicals among the three categories (very toxic: 2; moderately toxic: 1; and slightly/non-toxic: 0) of chemicals (Table 1). Accordingly, several ML-based classification models were built, and the best-performing classifiers are herein reported, which are decision trees (fine, medium and coarse) and Naïve Bayes. The model parameters and optimal architecture were determined by employing internal and external validation procedures. For internal validation, 5-fold cross-validation was employed, whereas, for external validation, a sub-set of the dataset, i.e., a test set (20 %), was used. The criteria to assess the predictive performance and reliability were set using sensitivity (*SE*), specificity (*SP*), area under curve (*AUC*) and model accuracy (*ACC*). The CV results (average of 10 repeats) for both classification models are summarised in Table 7.

The optimal DT model had the maximum number of splits as 100, 20 and 4, respectively, while the Gini's diversity index was employed as the split criterion. Each model had the *ACC*, *SE* and *SP* value of 100 % and *AUC* value of 1 for the 5-CV and test set, and as evident, performed the best for the classification of the three classes with no miscalculations. DT models, being non-parametric, do not make any assumptions about the distribution of the data. This makes them more flexible than parametric models like logistic regression, which assumes a linear relationship between the input features and the output (Abdalati et al., 2022).

In the case of optimal naïve Bayes, the model coupled with the Gaussian kernel performed reasonably well for the training set and performed better during the external validation. The model had the average *ACC*, *SE*, *SP* and *AUC* values of 91.5 %, 75.8 %, 96.4 % and 0.95, respectively; for 5-CV; and 94.4 %, 97.8 %, 96 % and 0.94, respectively, for the test set. During the 5-fold cross-validation process, the naïve Bayes model was able to classify highly toxic biocides with 100 % accuracy and no miscalculations, while 95 % accuracy during the classification of moderately toxic compounds with three miscalculations and 91.5 % accuracy during the classification of slightly/non-toxic compounds with three miscalculations. While during the external validation, the naïve Bayes model showed no miscalculations for the classification of moderately toxic and slightly/non-toxic biocides and only one miscalculation for the classification of highly toxic biocides. Naïve Bayes is, in general, a better classifier for similar tasks as it is robust to noise and irrelevant features because it assumes that features are

Table 7
Classification matrix for biocide toxicity prediction of 3-categories by different models.

Decision tree	Training set (5-fold Cross-Validation)						
	Actual class	Total instances	Predicted correct	Mis-classified	Model Accuracy (ACC)	SE (Sensitivity)	SP (Specificity)	AUC
	0	18	18	0	100 %	100 %	100 %	1
	1	12	12	0	100 %	100 %	100 %	1
	2	41	41	0	100 %	100 %	100 %	1
	Total	71						
	Test set (external validation)							
	Actual class	Total instances	Predicted correct	Mis-classified	Model Accuracy	SE (Sensitivity)	SP (Specificity)	AUC
	0	1	1	0	100 %	100 %	100 %	1
	1	1	1	0	100 %	100 %	100 %	1
	2	16	16	0	100 %	100 %	100 %	1
	Total	18						
Naïve Bayes	Training set (5-fold cross-validation)						
	Actual class	Total instances	Predicted correct	Mis-classified	Model Accuracy (ACC)	SE (Sensitivity)	SP (Specificity)	AUC
	0	18	15	3	91.5 %	83.3 %	94.3 %	0.96
	1	12	9	3	91.5 %	75.0 %	95.0 %	0.89
	2	41	41	0	91.5 %	69.4 %	100.0 %	1
	Total	71						
	Test set (external validation)							
	Actual class	Total instances	Predicted correct	Mis-classified	Model accuracy (ACC)	SE (Sensitivity)	SP (Specificity)	AUC
	0	1	1	0	94.4 %	100.0 %	94 %	0.94
	1	1	1	0	94.4 %	100.0 %	94 %	0.94
	2	16	15	1	94.4 %	93.5 %	100 %	0.94
	Total	18						

independent of each other. This means that even if some features are not relevant to the classification task or contain noise, the classifier can still perform well (Salmi and Rustam, 2019).

However, it is essential to note that the overall generalisability and reliability of such classifiers in the regulatory context rely on the predictive performance with comparatively large and balanced datasets, which was a limiting factor in this study. When evaluating the predictive performance of such models, it is also crucial to use appropriate metrics that accurately reflect the model's ability to predict the properties or activities of chemicals. Sensitivity, specificity, accuracy, and AUC can be less sensitive to class imbalance, but their performance can be affected by a class imbalance to some extent.

3.6. Applicability domain (AD) assessment

For reliable predictions, the applicability domain of the QSAR models was further analyzed using the software Applicability Domain v1.0 which follows the standardization approach to probe any outliers present in training and test set. According to this method, if the standardised value of a compound's molecular descriptors is $\leq$ 3, the compound is not an X-outlier or within AD, and vice versa. Only one compound in the test set was found to have an S_{new} value of 4.78, i.e., >3 (formaldehyde), suggesting an X-outlier or outside AD. While in the training set, four compounds had an S_{new} value of 3.15, 3.14, 5.33 and 3.28 (actane, dbnpa, neostanox and flubendiamide), implying X-outlier or outside the AD (appendix) (see Supplementary file 2). The outliers, nevertheless, were still incorporated during the model-building process due to fewer chemicals in the dataset, and the predictions were performed poorly for formaldehyde and neostanox only. This can be justified as only formaldehyde and neostanox had a considerably high S_{new} value, 4.78 and 5.33, respectively. A possible explanation for the detection of formaldehyde as an outlier in the test set is its relatively simple structure in comparison to the majority with highly diverse and complex structures. In addition, formaldehyde also had the lowest atomic LogP value (ALOGP), suggesting higher hydrophilicity and one hydrogen atom (H-049) directly attached to the carbon atom (C1) in formaldehyde, while one hydrogen atom (H-049) attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp) of another molecule. In the training set, neostanox had exceedingly high atomic LogP, suggesting a very high hydrophobic nature; this is due to the presence of non-polar functional groups, also resulting in high Atom-Type *E*-state (ATE). The relationship between ATE and logP is based on the fact that the electronic state of atoms in a molecule can influence the molecule's solubility and partitioning behaviour. In particular, atoms with higher ATE values (indicating a more electron-withdrawing or polar group) tend to be more hydrophilic and less likely to partition into non-polar solvents (Kier et al., 1999). In addition, neostanox was the only chemical with the presence of an [Sn] atom in the dataset. The presence of [Sn] molecular descriptor in the case of neostanox can significantly distinguish the substance from the dataset, eventually affecting the overall generalisability of the silico models. The other possible reason for the poor predictive performance of molecularly similar compounds could be factors such as erroneous, insufficient or poor-quality raw data used for training the model. Hence, it is recommended to exclude the detected outliers from the dataset in order to improve the overall generalisability and predictive performance of the model.

3.7. Adaptive modelling for reliable ecotoxicological evaluations in a regulatory context

The developed ML models presented in this report have shown good predictive performance, high generalisability, and the potential to replace animal testing for biocide ecotoxicological screening in marine crustaceans. However, its acceptance and the impact it merits in regulatory decision-making is still a topic of debate. The key arguments are (i) model generalisability and adaptability (ii) reliability of model validation (iii) confidence in predictive accuracy and (iv) transparency and interpretability of some ML algorithms. The OECD guidelines principle 2 provides important guidance on the quality and relevance of data used in chemical safety assessments. However, there are some limitations to its implementation, such as the limited availability of high-quality (LC_{50}) datasets for many chemicals. In some cases, there may be gaps in the data, or the available data may not be sufficient to fully characterize the risks associated with a chemical.

Principle 2 also emphasises “unambiguous algorithm”, which entails transparency and reproducibility of the models so that others can understand and reproduce the results. The intrinsic limitation to this is that some of the proposed models in this study, such as multi-layer feed-forward backpropagation ANN and other non-linear models, could be complex and might require technical expertise to understand and reproduce. Furthermore, ensuring transparency and reproducibility of models and algorithms used in chemical safety assessments requires significant resources, including time, expertise, and infrastructure. These resources may not always be available, particularly in the case of small and medium-sized enterprises or developing countries. A similar challenge also coincides with OECD guidelines principle 5 pertaining to the mechanistic interpretation of QSAR models. Biological systems are often complex and multifaceted, with many different pathways and interactions that can influence chemical activity. Mechanistic interpretation of such QSAR models may also oversimplify these systems, leading to inaccurate predictions.

Experimental validation is also an essential step in the development and evaluation of QSAR models for regulatory purposes. This validation process involves testing the model's predictions against experimental data to evaluate its accuracy and reliability (OECD, 2004). While experimental validation is certainly an important part of validating any scientific model or theory, it is not always feasible or necessary for QSAR models (Tropsha, 2010). This is because QSAR models are based on statistical relationships between chemical structures and biological activities. These relationships can be tested using various statistical measures, such as sensitivity, specificity, accuracy, precision, and the area under the receiver operating characteristic (ROC) curve (Grandini et al., 2020). These metrics provide information on the models' ability to correctly predict positive and negative cases and to distinguish between hazardous and non-hazardous chemicals. In addition, experimental validation can be time-consuming, costly, and sometimes unethical if it involves animal testing. QSAR models offer a faster, cheaper, and more ethical alternative to experimental testing. They can also be used to prioritise chemicals for further testing or to design new chemicals with specific properties, which can help to reduce the need for animal testing (Khan et al., 2019).

In our study, we employed k-fold cross-validation, where the entire dataset was divided into ten subsets, of which nine subset was used to train the model and the remaining subset was treated as a test set to validate the model. This method improves the robustness of the model to data variability by averaging the performance across multiple runs of the cross-validation process. This can help to reduce the impact of data variability on the model's predictive performance. A similar approach was adopted by Liu et al. (2019) to predict and validate chemical toxicity in marine crustaceans, where the classification models yielded fairly well results. Furthermore, for multi-class classification modelling, where the dataset is relatively small, and one class is more prevalent. It is important to use a combination of evaluation metrics, including those less sensitive to class imbalance. For example, Singh et al. (2013) employed a combination of sensitivity, specificity and accuracy, which measures the occurrence of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) in the multi-class classification of diverse chemicals acute toxicity in fish. A similar approach was also adopted by Liu et al. (2019) to classify acute chemical toxicity in marine crustaceans. Various other multi-class classification evaluation metrics such as Matthews Correlation Coefficient (*MCC*), Cohen's Kappa, macro-averaged precision, recall, and F1-score can also provide a more accurate assessment of the model's predictive performance in the presence of class imbalance (Grandini et al., 2020).

3.8. Comparison of developed models with models available in the literature

The LC_{50} is a widely used endpoint in QSAR modelling, particularly in the field of ecotoxicology. Such QSAR models that predict LC_{50} values

can provide valuable information for regulatory decision-making and environmental risk assessment (ERA). However, the literature survey showed that the potential of computational models to predict biocide LC_{50} in marine crustaceans had not yet been extensively explored. Therefore, a quantitative comparison with others' work would be irrelevant because the datasets and target organisms differ between the models. Nonetheless, a simple comparison of our model methodology and result statistics will give fundamental insight into the accuracy of various approaches to building such models.

Various classification-based models were developed by Liu et al. (2019) to predict and classify the LC_{50} values of a wide array of chemicals in marine crustaceans. The method employed six ML models, which are SVM, NB, RF, DT, kNN, and ANN, and trained using a set of 1D/2D molecular descriptors and fingerprints. Similar 10-fold cross-validation was also employed for model validation, and the *AUC* values of the developed models ranged from 0.80 to 0.90 for test sets. The DT model developed in our study showed the *AUC* value of 1 for both the training and test set. However, it is important to note that the models developed by Liu et al. (2019) used a significantly large dataset (>1000) which was a limiting factor in our study. For the acceptance of a model in a regulatory context, it is also recommended that the models are trained using a large and good-quality dataset. Similarly, two partial least squares (PLS) regression-based models were developed by Khan et al. (2019) to predict LC_{50} values of biocides in *Daphnia magna* and fish toxicities using 2D descriptors. The method employed leave-one-out cross-validation to validate the models, and the results yielded R^2 of 0.80 and 0.64, respectively, for fish training and test set, and R^2 0.87 and 0.81, respectively, for *Daphnia magna* training and test set. These models showed satisfactory results; however, they tend to overfit the training set. Overfitting occurs when a model learns the patterns in the training data too well and becomes too specific to that data. As a result, the model may not generalize well to new, unseen data, such as the test set. The presented models in our study have shown high generalisability by avoiding overfitting on the training data suggesting appropriateness to replace unnecessary animal testing to predict biocide toxicity in a wide range of marine crustacean species.

4. Conclusions

In this study, firstly, an overview was presented on how extensive use of biocidal products can have a detrimental impact on the aquatic organisms, with particular reference to crustaceans due to their non-target mechanism of action. Secondly, in the light of incorporating animal alternatives for environmental risk assessment (ERA) of hazardous chemicals, *in silico* models were built to fill this data gap by predicting the acute chemical toxicity of biocidal chemicals in environmentally sensitive invertebrates - marine crustaceans. The work presented herein has shown that *in silico* modelling approaches are a powerful method to predict acute chemical toxicity of biocides, enabling rapid prioritisation of compounds during ERA. The biocide dataset used in the research shows good diversity, and each predictive model is quite varied in its approach as well. All six models in this study yielded satisfactory results, and the feed-forward backpropagation-based artificial neural network model showed the best performance during regression analysis, while decision tree model performed the best for the classification of different toxicities. Nevertheless, ML approaches have great potential in ecotoxicological studies, and further improvement and understanding of the underlying science are important. The major limiting factor in this study to build an even more robust model was the small biocide sample size of the dataset ($n = 89$); hence, updating the chemical and ecotoxicological databases is also pivotal. In addition to predicting the toxicity of a particular chemical, ML can also be used to interpret the influence of a particular molecular descriptor or property contributing to its toxicity, allowing to manufacture of a greener and more sustainable chemical product. The developed models are capable of predicting the toxicities of untested biocides within the applicability domain of the models.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.164072.

CRediT authorship contribution statement

RK: Data curation, Formal analysis, Methodology, Software, Validation, Writing original draft, Review and editing.

ISH: Methodology, Software, Validation, Supervision, Review and editing.

SC: Data curation, Formal analysis, Methodology, Validation.

ANJ: Conceptualization, Methodology, Validation, Supervision, Resources, Project administration, Review and editing.

Data availability

Data will be made available on request.

Declaration of competing interest

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.

Acknowledgement

The work has been carried out as a part of Master's in Research (MRes) degree programme at the University of Plymouth, UK. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abdalati, A., Saed, A., Jaharadak, A.A., 2022. Implementation with performance evaluation of decision tree classifier for uncertain data: literature review. Int. J. Multidiscip. Res. Publ. 5 (5), 125–132.

Altman, D.G., Bland, J.M., 1994. Statistics Notes: diagnostic tests 2: predictive values. Bmj 309 (6947), 102.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054.

Backhaus, T., Altenburger, R., Faust, M., Frein, D., Frische, T., Johansson, P., Kehrer, A., Porsbring, T., 2013. Proposal for environmental mixture risk assessment in the context of the biocidal product authorization in the EU. Environ. Sci. Eur. 25 (1), 1–9. https://doi.org/10.1186/2190-4715-25-4/FIGURES/2.

Barnston, A.G., 1992. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. Weather Forecast. 7 (4), 699–709.

Barros, R.P.C., Sousa, N.F., Scotti, L., Scotti, M.T., 2020. Use of machine learning and classical QSAR methods in computational ecotoxicology. Methods Pharmacol. Toxicol. 151–175. https://doi.org/10.1007/978-1-0716-0150-1_7.

Berthold, M.R., Cebron, N., Dill, F., Di Fatta, G., Gabriel, T.R., Georg, F., Meinl, T., Ohl, P., Sieb, C., Wiswedel, B., 2009. KNIME - the Konstanz information miner. ACM SIGKDD Explorations Newsletter, pp. 58–61 https://doi.org/10.1145/1656274.1656280.

Bickel, P.J., Doksum, K.A., 2015. Mathematical Statistics: Basic Ideas and Selected Topics, Volume I. CRC Press.

Bilski, J., Kowalczyk, B., Marchlewska, A., Zurada, J.M., 2020. Local levenberg-marquardt algorithm for learning feedforwad neural networks. JAISCR 10 (4), 299. https://doi.org/10.2478/jaiscr-2020-0020.

Chandrasekaran, B., Abed, S.N., Al-Attraqchi, O., Kuche, K., Tekade, R.K., 2018. Computer-aided prediction of pharmacokinetic (ADMET) properties. Dosage Form Design Parameters. 2, pp. 731–755. https://doi.org/10.1016/B978-0-12-814421-3.00021-X.

Chang, Y.W., Hsieh, C.J., Chang, K.W., Ringgaard, M., Lin, C.J., 2010. Training and testing low-degree polynomial data mappings via linear SVM. J. Mach. Learn. Res. 11 (4).

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–17-August-2016, pp. 785–794 https://doi.org/10.1145/2939672.2939785.

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, 1–13.

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2013. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Routledge.

COMBASE, 2016. COMBASE. https://www.life-combase.com/index.php/en/.

Coors, A., Vollmar, P., Heim, J., Sacher, F., Kehrer, A., 2018. Environmental risk assessment of biocidal products: identification of relevant components and reliability of a component-based mixture assessment. Environ. Sci. Eur. 30 (1), 1–15. https://doi.org/10.1186/S12302-017-0130-0/TABLES/4.

Dale, V.H., Beyeler, S.C., 2001. Challenges in the development and use of ecological indicators. Ecol. Indic. 1 (1), 3–10. https://doi.org/10.1016/S1470-160X(01)00003-6.

Damodar, N.G., Dawn, C.P., 2009. Basic econometric fifth edition. McGraw-Hill.

Demšar, J., Erjavec, A., Hočevar, T., Milutinovič, M., Možina, M., Toplak, M., Umek, L., Zbontar, J., Zupan, B., 2013. Orange: data mining toolbox in Python Tomaž Curk Matija Polajnar Laň Zagar. J. Mach. Learn. Res. 14, 2349–2353.

Devi, T.G., Patil, N., Rai, S., Sarah, C.P., 2023. Segmentation and classification of white blood cancer cells from bone marrow microscopic images using duplet-convolutional neural

network design. Multimed. Tools Appl. 1–23. https://doi.org/10.1007/S11042-023-14899-9/METRICS.
Ebenuwa, S.H., Sharif, M.S., Alazab, M., Al-Nemrat, A., 2019. Variance ranking attributes selection techniques for binary classification problem in imbalance data. IEEE Access 7, 24649–24666. https://doi.org/10.1109/ACCESS.2019.2899578.
EC, 2009. Assessment of different options to address risks from the use phase of biocides. Final Report . www.cowi.com.
EC, 2018. Report from the commission to the European parliament and the council on the implementation of the Union authorisation of biocidal products in accordance with Article 42(3) of Regulation (EU) No 528/2012 of the European Parliament and of the Council concerning the making available on the market and use of biocidal products. https://ec.europa.eu/health/biocides/regulation_en.
ECHA, 2022. Homepage - *ECHA*. https://echa.europa.eu/.
EU, 2012. Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 Concerning the Making Available on the Market and use of Biocidal Products. ISSN 1977 677p. 2985.
Flemming, H.-C., Murthy, P.S., Venkatesan, R., Cooksey, K., 2009. Marine and Industrial Biofouling. vol. 333. Springer.
Gini, C., 1936. On the measure of concentration with espacial reference to income and wealth. Cowles Comm. 2 (3).
Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: an overview. https://arxiv.org/abs/2008.05756v1.
Han, H., Jiang, X., 2014. Overcome support vector machine diagnosis overfitting. Cancer Informat. 13, CIN–S13875.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.
Hansch, C., Fujita, T., 1964. ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc. 86 (8), 1616–1626. https://doi.org/10.1021/JA01062A035/ASSET/JA01062A035.FP.PNG_V03.
Ho, T.K., 1995. Random decision forests. Proceedings of the International Conference on Document Analysis and Recognition. 1. ICDAR, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.
Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S.A., Attene-Ramos, M., Zhao, T., Austin, C.P., Simeonov, A., 2016. Modelling the Tox21 10K chemical profiles for toxicity prediction and mechanism characterization. Nat. Commun. 7 (1), 1–10. https://doi.org/10.1038/ncomms10425.
Isabona, J., Imoize, A.L., Kim, Y., 2022. Machine learning-based boosted regression ensemble combined with hyperparameter tuning for optimal adaptive learning. Sensors 22 (10), 3776. https://doi.org/10.3390/S22103776.
Kensert, A., Alvarsson, J., Norinder, U., Spjuth, O., 2018. Evaluating parameters for ligand-based modeling with random forest on sparse data sets. J. Cheminformatics 10 (1), 49. https://doi.org/10.1186/S13321-018-0304-9.
Khan, K., Khan, P.M., Lavado, G., Valsecchi, C., Pasqualini, J., Baderna, D., Marzo, M., Lombardo, A., Roy, K., Benfenati, E., 2019. QSAR modeling of *Daphnia magna* and fish toxicities of biocides using 2D descriptors. Chemosphere 229, 8–17. https://doi.org/10.1016/J.CHEMOSPHERE.2019.04.204.
Kier, L.B., Hall, L.H., 1937, 1999. Molecular Structure Description. 41 doi:10.3/JQUERY-UI.JS.
Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E., 2021. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 49 (D1), D1388–D1395. https://doi.org/10.1093/NAR/GKAA971.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26. https://doi.org/10.18637/JSS.V028.I05.
Langdon, C.J., Vance, M.M., Harmon, V.L., Kreeger, K.E., Kreeger, D.A., Chapman, G.A., 1996. A 7-D toxicity test for marine pollutants using the pacific mysid *Mysidopsis intii*. 1. Culture and protocol development. Environ. Toxicol. Chem. 15 (10), 1815–1823. https://doi.org/10.1002/ETC.5620151024.
Liu, R., Madore, M., Glover, K.P., Feasel, M.G., Wallqvist, A., 2018. Assessing deep and shallow learning methods for quantitative prediction of acute chemical toxicity. Toxicol. Sci. 164 (2), 512–526. https://doi.org/10.1093/toxsci/kfy111.
Liu, L., Yang, H., Cai, Y., Cao, Q., Sun, L., Wang, Z., Li, W., Liu, G., Lee, P.W., Tang, Y., 2019. *In silico* prediction of chemical aquatic toxicity for marine crustaceans via machine learning. Toxicol. Res. 8 (3), 341–352. https://doi.org/10.1039/c8tx00331a.
Lussier, S.M., Kuhn, A., Comeleo, R., 1999. An evaluation of the seven-day toxicity test with *Americamysis bahia* (formerly *Mysidopsis bahia*). Environ. Toxicol. Chem. 18 (12), 2888–2893. https://doi.org/10.1002/ETC.5620181233.
Marzo, M., Lavado, G.J., Como, F., Toropova, A.P., Toropov, A.A., Baderna, D., Cappelli, C., Lombardo, A., Toma, C., Blázquez, M., Benfenati, E., 2020. QSAR models for biocides: the example of the prediction of *Daphnia magna* acute toxicity. SAR QSAR Environ. Res. 31 (3), 227–243. https://doi.org/10.1080/1062936X.2019.1709221/SUPPL_FILE/GSAR_A_1709221_SM4833.DOCX.
MATLAB, 2010. Version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts.
Mauri, A., Srl, A., 2021. Development of Software Tools for the Application of QSAR Models View Project OpenTox View Project Chapter 32 alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. https://doi.org/10.1007/978-1-0716-0150-1_32.
Miller, T.H., Gallidabino, M.D., Macrae, J.I., Hogstrand, C., Bury, N.R., Barron, L.P., Snape, J.R., Owen, S.F., 2018. Machine learning for environmental toxicology: a call for integration and innovation. Environ. Sci. Technol. 52 (22), 12953–12955 American Chemical Society https://doi.org/10.1021/acs.est.8b05382 American Chemical Society.
Miller, T.H., Gallidabino, M.D., MacRae, J.R., Owen, S.F., Bury, N.R., Barron, L.P., 2019. Prediction of bioconcentration factors in fish and invertebrates using machine learning. Sci. Total Environ. 648, 80–89. https://doi.org/10.1016/j.scitotenv.2018.08.122.
Oberdörster, E., Cheek, A.O., 2001. Gender benders at the beach: endocrine disruption in marine and estuarine organisms. Environ. Toxicol. Chem. 20 (1), 23–36. https://doi.org/10.1002/ETC.5620200103.
OECD, 2004. Validation of (Q)SAR models. https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm.
Olker, J.H., Elonen, C.M., Pilli, A., Anderson, A., Kinziger, B., Erickson, S., Skopinski, M., Pomplun, A., LaLone, C.A., Russom, C.L., Hoff, D., 2022. The ECOTOXicology knowledgebase: a curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. Environ. Toxicol. Chem. 41 (6), 1520–1539. https://doi.org/10.1002/ETC.5324.
Rand, G.M., 1985. Introduction. In: Rand, G.M., Petrocelli, S.R. (Eds.), Fundamentals of Aquatic Toxicology: Methods and Application. 1. Hemisphere Publishing Corporation. Cap, London, pp. 1–28.
Rennie, J.D., Shih, L., Teevan, J., Karger, D.R., 2003. Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the 20th international conference on machine learning (ICML-03), pp. 616–623.
Roast, S.D., Thompson, R.S., Donkin, P., Widdows, J., Jones, M.B., 1999. Toxicity of the organophosphate pesticides chlorpyrifos and dimethoate to *Neomysis integer* (Crustacea: *Mysidacea*). Water Res. 33 (2), 319–326. https://doi.org/10.1016/S0043-1354(98)00248-6.
Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50 (5), 742–754. https://doi.org/10.1021/CI100050T/ASSET/IMAGES/MEDIUM/CI-2010-00050T_0018.GIF.
Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. 145, 22–29. https://doi.org/10.1016/J.CHEMOLAB.2015.04.013.
Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E., Drummond, R.A., 1997. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). Environ. Toxicol. Chem. 16 (5), 948–967. https://doi.org/10.1002/ETC.5620160514.
Salmi, N., Rustam, Z., 2019. Naïve Bayes classifier models for predicting the colon cancer. IOP Conf. Ser.: Mater. Sci. Eng. 546 (5), 052068. https://doi.org/10.1088/1757-899X/546/5/052068.
Schüürmann, G., Ebert, R.U., Kühne, R., 2011. Quantitative read-across for predicting the acute fish toxicity of organic compounds. Environ. Sci. Technol. 45 (10), 4616–4622. https://doi.org/10.1021/ES200361R.
Sieg, J., Flachsenberg, F., Rarey, M., 2019. In need of Bias control: evaluating chemical data for machine learning in structure-based virtual screening. J. Chem. Inf. Model. 59 (3), 947–961. https://doi.org/10.1021/ACS.JCIM.8B00712/SUPPL_FILE/CI8B00712_SI_001.PDF.
Singh, K.P., Gupta, S., Rai, P., 2013. Predicting acute aquatic toxicity of structurally diverse chemicals in fish using artificial intelligence approaches. Ecotoxicol. Environ. Saf. 95, 221–233. https://doi.org/10.1016/j.ecoenv.2013.05.017.
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., ... Tetko, I.V., 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput. Aided Mol. Des. 25 (6), 533–554. https://doi.org/10.1007/S10822-011-9440-2.
Tahmasebi, P., Hezarkhani, A., 2011. Application of a modular feedforward neural network for grade estimation. Nat. Resour. Res. 20, 25–32.
Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Mol. Inform. 29 (6–7), 476–488. https://doi.org/10.1002/MINF.201000061.
US-EPA, 2021. Technical overview of ecological risk assessment - analysis phase: ecological effects characterization | US EPA. https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/technical-overview-ecological-risk-assessment-0.
Walters, P., 2019. Visualizing chemical space. November 1 http://practicalcheminformatics.blogspot.com/2019/11/visualizing-chemical-space.html.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.
Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., Yu, Y., 2018. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. http://arxiv.org/abs/1810.13306.
Yazdi, M., Golilarz, N.A., Nedjati, A., Adesina, K.A., 2021. An improved lasso regression model for evaluating the efficiency of intervention actions in a system reliability analysis. Neural. Comput. Appl. 33 (13), 7913–7928. https://doi.org/10.1007/S00521-020-05537-8.