

**University of Plymouth**

**PEARL**

**<https://pearl.plymouth.ac.uk>**

---

Faculty of Science and Engineering

School of Biological and Marine Sciences

---

2023-08-20

# ***In Silico* Prediction of Acute Chemical Toxicity of Biocides in Marine Crustaceans Using Machine Learning**

Rama Krishnan <sup>a</sup>, Ian S. Howard <sup>b</sup>, Sean Comber <sup>c</sup>, Awadhesh N. Jha <sup>a\*</sup>

<sup>a</sup>*School of Biological and Marine Sciences, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK.*

<sup>b</sup>*School of Engineering, Computing and Mathematics, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK.*

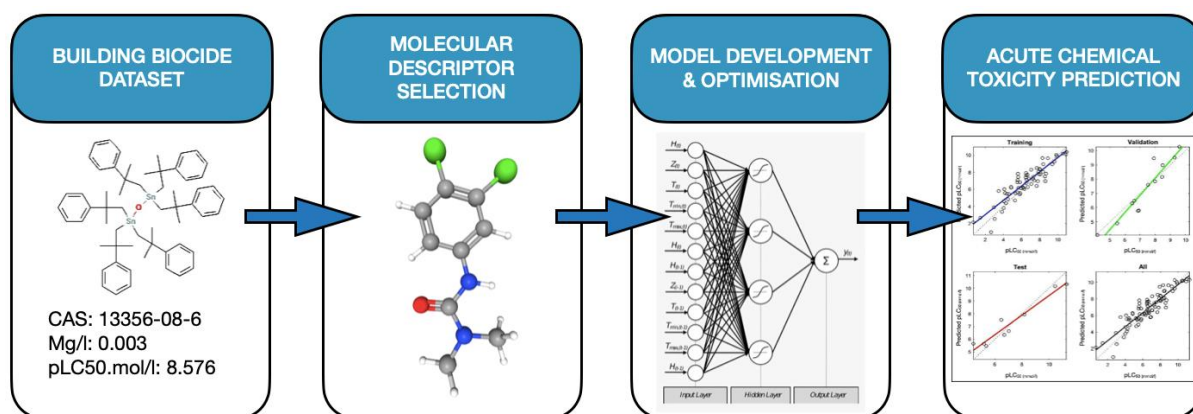
<sup>c</sup>*School of Geography, Earth and Environmental Sciences, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK.*

\*Corresponding author E-mail address: a.jha@plymouth.ac.uk (Awadhesh N. Jha)  
School of Biological and Marine 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

## **Krishnan et al. Graphical Abstract**



## 1 **Abstract**

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

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

28 **Keywords:** Biocides, LC<sub>50</sub>, Machine Learning, QSAR, Marine Crustaceans,  
29 Ecotoxicology

## 30 **1. Introduction**

31 Biocides are a heterogeneous group of chemicals which are used “with the  
32 intention of destroying, deterring, rendering harmless, preventing the action of, or  
33 otherwise exerting a controlling effect on, any harmful organism by any means other  
34 than mere physical or mechanical action” (EU, 2012). These biocides comprise of an  
35 “active substance” incorporated with “co-formulants” (such as stabilizers, solvents,  
36 carriers and wetting agents) to ensure the final potency of biocidal mixture (Marzo et  
37 al., 2020). These biocidal products, *via* point and non-point sources, enter the aquatic  
38 environments and may pose a threat to ecologically and commercially important non-  
39 target organisms with long-term impact on the ecosystems, and human health (Coors  
40 et al., 2018; Flemming et al., 2009). For example, in Europe, biocidal products are  
41 regulated by the BPR, Regulation (EU: 528/2012) (EU, 2012). According to the current  
42 biocidal product regulation (EU, 2012), the formulation, including both “active  
43 substance” and “co-formulants”, must undergo an environmental risk assessment  
44 (ERA) to evaluate the toxicity of biocidal products (Backhaus et al., 2013). Moreover,  
45 this regulation improves the efficiency of internal market harmonizing rules and  
46 ensures effective protection of the animals and human health and the environment.  
47 Additionally, the European Chemicals Agency (ECHA) also ensures the overall  
48 applicability and robustness of the legislation by providing technical and scientific  
49 support to the European Commission (EC) (EC, 2018). The biocides can be classified

50 into 22 product types (PT) (Marzo et al., 2020) which are further categorized into four  
51 groups (Khan et al., 2019). The active substance specific to the PTs also determines  
52 their approval.

53         There are official risk assessment reports by the EC addressing various  
54 ecotoxicological risks caused by the use of specific PTs (EC, 2009). The reports  
55 suggest that the biocides can be carried away to non-target sites during their  
56 applications (e.g., during rain via runoff), including the surface water, signifying a threat  
57 to the aquatic ecosystem. Sustainable use of biocides is therefore imperative. It is also  
58 necessary to emphasize the need to understand the short and long-term  
59 consequences of biocides on the aquatic ecosystem and the valuable resources  
60 therein. Consequently, in 2016, the EC initiated the LIFE-COMBASE project  
61 (COMBASE, 2016). The project aims to promote and encourage the sustainable use  
62 of biocides by analyzing the overall risks they pose to the environment and human  
63 health. The LIFE-COMBASE project also promotes chemical hazard assessment  
64 using alternative methods to animal testing by incorporating *in silico* approaches. The  
65 introduction of an innovative approach for environmental health monitoring using the  
66 application of machine learning (ML) has recently attracted attention in  
67 ecotoxicological studies. The implementation of ML in this context is based on the use  
68 of algorithms allowing the system to learn, interpret, and predict the chemical and  
69 biological processes associated with it (Miller et al., 2018). With the advancement in  
70 these computational approaches, such as read-across (RA) and quantitative structure-  
71 activity relationships (QSARs), ML facilitates efficient risk management by eliminating  
72 and outperforming unnecessary testing on animals while less time-consuming  
73 concurrently (Liu et al., 2018; Miller et al., 2018). A plethora of studies is available  
74 reporting that ML approaches in QSAR surpass other computation-based

75 conventional approaches, for instance, knowledge-based functions of datasets and  
76 empirical scoring methodologies (Sieg et al., 2019; Barros et al., 2020). Nevertheless,  
77 understanding the underlying science and rationale behind selecting features,  
78 algorithms and interpretation knowledge is crucial (Sieg et al., 2019; Barros et al.,  
79 2020).

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

92 In the backdrop of above information, our study aimed to build highly predictive  
93 and robust *in silico* models. These models were validated through stringent processes  
94 to probe the acute chemical toxicity of various biocides on marine crustaceans. In  
95 order to achieve the objectives, firstly, an acute chemical toxicity or LC<sub>50</sub> dataset was  
96 built, which is the mean lethal concentration, determining the concentration of a  
97 substance in the medium causing mortality to 50% of a group of test organisms within  
98 a period of exposure (Rand, 1985). The toxicity data were generated for the three

99 families of marine crustaceans, including *Mysidae*, *Palaemonidae* and *Penaeidae*.  
100 Subsequently, regression and classification-based computational models were built to  
101 predict the biocide toxicity in these marine crustaceans. In predictive models, the  
102 chemicals were represented as molecular descriptors. Following this, the key  
103 molecular descriptors influencing acute chemical toxicity were investigated using ML  
104 methods. The molecular descriptors were also employed to check the applicability  
105 domain of the chemicals in the dataset.

## 106 **2. Materials and methods**

### 107 *2.1. Dataset Sources*

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

## 121 2.2. Dataset pre-processing

122 For modelling purposes and to improve the overall performance of ML models, the  
123 compounds with incorrect CAS numbers or molecular structures not clearly identified  
124 were removed from the dataset. Furthermore, to retain an uniformity of biocides in the  
125 dataset, metal complexes, inorganic compounds, mixtures with unknown  
126 compositions, and salts containing organic counterions were removed. Additionally,  
127 the structure of the remaining salts in the dataset was also neutralized. From the  
128 dataset containing biocides to be used for modelling, all LC<sub>50</sub> units were first converted  
129 to parts per million (ppm) and data with units that could not be directly converted, for  
130 example, AI (active ingredient) ppm, AI µg/l, and mol/l were removed. Later, the  
131 duplicates were removed, and the geometric mean of similar compounds with multiple  
132 experimental values was calculated. Finally, the observed values expressed as ppm  
133 (or mg/l) were converted to mmol/l followed by negative logarithmic transformation (-  
134 Log 10 mmol/l) or p-transformation, i.e., pLC<sub>50</sub>, in accordance with ecotoxicological  
135 QSAR studies. The purpose of p-transformation is to reduce the skewness of the data,  
136 which can be beneficial for statistical analysis that assume normally distributed data.  
137 Consequently, higher pLC<sub>50</sub> values corresponded to higher toxicity and *vice versa*.

138 For classification modelling, the guidelines provided by the US-EPA were followed,  
139 which suggests classifying the different toxicity categories of chemicals for ecological  
140 risk assessment. Accordingly, the chemical aquatic toxicity (ppm) can be classified  
141 into five categories, i.e., very highly toxic (<0.1), highly toxic (0.1-1), moderately toxic  
142 (>1-10), slightly toxic (>10-100), and non-toxic (>100) (US-EPA, 2021).



### 143 2.3. *Calculation of molecular descriptor*

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

152 To avoid any conformational complexity and for ease of interpretability, only 2D  
153 molecular descriptors were calculated in this study. These molecular descriptors were  
154 retrieved from the 2D characterization of molecular structures, which quantify the  
155 molecular characteristics such as connectivity of atoms in a molecule and atomic  
156 composition (Mauri & Srl, 2021). Firstly, the SMILE strings for each molecule were  
157 created, which are the linear structural concepts describing the structure of chemical  
158 species. Secondly, in total, 2223 molecular descriptors were calculated, comprising of  
159 2D atom pairs, atom type E-state indices, functional group counts, constitutional  
160 indices, topological indices, ring descriptors, atom-centred fragment molecular  
161 property, and 2D molecular descriptors were calculated using PaDEL2 and Dragon v.  
162 7 from the open access OCHEM database (Sushko et al., 2011). Additionally, the  
163 RDKit 2D molecular descriptors were also calculated using KNIME Analytics Platform  
164 version 4.3.1 (Berthold et al., 2009).

#### 165 2.4. Feature selection and dataset division

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

#### 180 2.5. Diversity in dataset

181 To develop a robust model with high accuracy and reliable predictions, it is crucial  
182 that the chemicals in the dataset are diverse. The diversity of chemicals in our dataset  
183 was investigated by first calculating Morgan (2D circular) fingerprints of radius 2 and  
184 1024 nBits for each chemical. The rationale behind selecting the specific fingerprint  
185 can be found in previous studies (Kensert et al., 2018; Liu et al., 2019). Secondly, the  
186 Tanimoto similarity index was calculated, which can be explained by the equation:  
187  $S_{A,B} = c/[a + b - c]$  and  $S = 1/(1 + \text{distance})$ , where S denotes similarities, a and b  
188 represent the number of bits in molecule A and B, respectively; while c represents the

189 number of bits that are in both molecules. Lastly, a heatmap was created to compare  
190 the similarities of each chemical. The entire process was performed using KNIME v  
191 4.3.1 (Berthold et al., 2009). In addition, principal component analysis (PCA) was also  
192 implemented to define the chemical space occupied by the compounds and diversity  
193 in the dataset. The PCA analysis takes the high-dimensional sets of correlated  
194 molecular properties or molecular descriptors into consideration and combines them  
195 to create a lower-dimensional space of the corresponding properties making it easier  
196 to illustrate and interpret the molecular diversity (Walters, 2019).

## 197 2.6. Model building

198 For regression models, four supervised ML algorithms were employed, which are  
199 random forest (RF), artificial neural network (ANN), linear regression (LR), and support  
200 vector machine (SVM). In supervised learning, the algorithm is trained using “labelled”  
201 datasets and the prediction/classification is based on the data provided (Yao et al.,  
202 2018).

203 The SVM, LR and RF algorithms were implemented in Orange v 3.26.0 (Demšar  
204 et al., 2013), and the dataset was split into subsets so that 62 compounds (80%) were  
205 used to train the model (training set) and 17 compounds (20%) were used to test the  
206 model (test set). In the case of ANN, feed-forward backpropagation method was  
207 employed using Neural Net Fitting app in MATLAB R2021a (MATLAB, 2010) and the  
208 model was trained using the Levenberg-Marquardt technique. The dataset was split  
209 into 67 compounds (75%) as a training set, 13 compounds (15%) as validation set and  
210 9 compounds (10%) as test set. The ANN model consisted of one input layer with ten  
211 neurons (number of features), one hidden layer consisting of seven neurons  
212 (iteratively tuned and configured for best performance) and one output layer consisting

213 of one neuron. The Tan-Sigmoid transfer function (*tansig*) was employed in the hidden  
214 layer, while for the output layer, the Linear Transfer function (*purelin*) was employed.  
215 The architecture used to build the ANN model is illustrated in Fig 1. Similarly, for  
216 classification modelling, two supervised ML algorithms were employed, which are  
217 decision tree (DT) and naïve Bayes (NB). These algorithms were implemented in  
218 MATLAB R2021a (MATLAB, 2010). The details of these ML algorithms and  
219 configurations are mentioned in Table 1. More theoretical and mathematical details  
220 can be found in previous studies (Liu et al., 2019; Miller et al., 2019; Russom et al.,  
221 1997; Schüürmann et al., 2011; Singh et al., 2013).

#### 222 *2.6.1. Validation and performance evaluation*

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

## 238 2.6.2. Applicability domain (AD) study

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

$$245 S_{ki} = \frac{|X_{ki} - \bar{X}_i|}{\sigma_{X_i}}$$

246 Where  $k$ = total no. of compounds,  $i$ = total no. of descriptors,  $S_{ki}$ = standardised  
247 descriptors,  $X_{ki}$ = original descriptors,  $\bar{X}_i$ = mean of  $X_{ki}$ ,  $\sigma_{X_i}$ = standard deviation of  $X_{ki}$   
248 for training set.

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

$$253 S_{new(k)} = \bar{S}_k + 1.28 \times \sigma_{S_k}$$

254 Where,  $S_{new(k)}$  =  $S_{new}$  value for compound  $k$ ,  $\bar{S}_k$ = mean of  $S_{i(k)}$ ,  $\sigma_{S_k}$  = standard  
255 deviation of  $S_{i(k)}$ .

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

## 257 **3. Results and discussion**

### 258 *3.1. Dataset analysis*

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

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

### 281 3.2. *Diversity analysis in dataset*

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

293 Additionally, the Morgan (2D circular) fingerprints of radius 2 and 1024 nBits were  
294 used to construct a Tanimoto similarity heatmap which defined the similarity matrix for  
295 each compound (Fig. 2d), where the similarity increased from zero (blue) to one (red).  
296 Morgan fingerprints are a type of circular fingerprint that encode molecular structure  
297 information as a bit string. They are particularly useful for measuring diversity in a  
298 dataset since they capture important structural features of molecules relevant to their  
299 biological activity (Rogers & Hahn, 2010). The heatmap revealed that the substance  
300 in our dataset was diverse. Overall, the figures (Fig. 2 a-d) illustrate a good diversity  
301 of chemicals throughout the dataset.

302 3.3. *Molecular descriptor feature selection and relevance to toxicity*  
303 *prediction*

304 In conjunction with the quality of dataset used, selecting the most relevant  
305 molecular descriptors for toxicity prediction is crucial for optimizing the models and  
306 unravelling the molecular factors contributing to toxicity. To improve the overall  
307 generalisability and to avoid overfitting in our QSAR models, feature selection of the  
308 initially calculated molecular descriptors was performed. The features from the initial  
309 pool of 2223 molecular descriptors retrieved from Dragon v. 7, PaDEL 2 and RDKit  
310 were reduced using feature selection techniques such as *nearZeroVar*,  
311 *findCorrelation*, XGBoost and Best Subset Selection (BSS). From the initial pool of  
312 2223 molecular descriptors, 1825 molecular descriptors having more than 80% zero  
313 values and inter-correlated features ( $>0.90$ ) were eliminated from the dataset using  
314 *nearZeroVar* and *findCorrelation* function in RStudio. From the remaining 398  
315 molecular descriptors, the top 20 were reserved using XGBoost regression modelling  
316 in python3, and finally, the top 10 molecular descriptors were selected using the best  
317 subset selection (BSS) and used in regression modelling, which are: VE1\_Dt,  
318 VE2\_Dt, B07[C-C], H.049, C.002, ALOGP, XLogP, MLFER\_S, SRW10 and SMR.  
319 While for classification, eighteen descriptors were selected and used by employing  
320 XGBoost classification approach in python3 to build the final classification models,  
321 which are: Psi\_e\_1, nRCN, H.049, F01.C.N., F05.N.O., TPSA.NO., ALogP, ATSC1c,  
322 ATSC0p, MATS1v, MATS4p, GATS1i, MIC5, JGI6, Chi3v, Chi4v, slogp\_VSA10 and  
323 smr\_VSA3. The XGBoost feature selection for classification modelling works by  
324 selecting the most important features and can reduce the noise in the data, making it  
325 easier for the algorithm to find meaningful patterns. This often leads to improved model



326 performance, as the algorithm can focus on the most relevant features for the  
327 classification task (Devi et al., 2023).

328 Additionally, to assess the relevancy of the selected molecular descriptors to  
329 predict toxicity, the Pearson correlation ( $r$ ) method was employed for the set of  
330 molecular descriptors in regression analysis. This method is commonly used to  
331 measure the linear relationship between two continuous variables, where the  $r$ -value  
332 ranges from -1 to 1, with -1 indicating a perfectly negative linear relationship, 0  
333 indicating no linear relationship, and 1 indicating a perfectly positive linear relationship  
334 (Ebenuwa et al., 2019). The  $r$ -values of the features used for regression were retrieved  
335 in the order: ALOGP: +0.703; SRW10: +0.606; SMR: +0.603; VE1\_Dt: +0.599;  
336 XLogP: +0.578; MLFER\_S: +0.410; VE1\_Dt: +0.373; H.049: -0.222; C.022: -0.031.

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

346 It is important to note that while the Pearson correlation method is widely used to  
347 measure the relevancy of the features, it does have some limitations. Firstly, it only  
348 captures linear relationships between variables, meaning it may miss important non-  
349 linear relationships. Secondly, it only measures the relationship between two variables

350 at a time, and may not account for the effects of multiple variables on the target  
351 variable. To address these limitations, researchers can use more advanced  
352 techniques, such as regularisation methods like Lasso or Ridge regression, which can  
353 capture non-linear relationships and account for multiple variables simultaneously.

354 In addition to ALOGP, VE1\_Dt and VE2\_Dt are molecular descriptors that measure  
355 the topological complexity of a molecule. In general, molecules with higher values of  
356 VE1\_Dt and VE2\_Dt tend to be more hydrophobic and less soluble in water, while  
357 molecules with lower values tend to be more hydrophilic and more soluble. BO7[C-C]  
358 calculates the number of pairs of carbon atoms separated by a distance of 7 or fewer  
359 bonds. MLFER\_S is a useful molecular descriptor for predicting the solubility of drugs  
360 and other bioactive molecules, as solubility is a key factor affecting a drug's  
361 bioavailability and pharmacokinetics (Huang et al., 2016). SRW10 is a type of  
362 topological descriptor that represents the presence and distribution of various  
363 substructures within a molecule. It is useful for QSAR modelling in particular as it  
364 captures information about specific substructures that may be important for binding to  
365 the target (Hansch & Fujita, 1964).

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

### 372 3.4. Regression modelling

373 The regression models to predict the acute toxicity (pLC<sub>50</sub>) of biocide chemicals  
374 were built using our four best-performing modelling approaches (RF, SVM, LR, ANN).  
375 The overall generalisability, robustness and predictive performance were determined  
376 through stringent internal and external validation procedures. For internal validation,  
377 10-fold cross-validation was employed, whereas, for external validation, a sub-set of  
378 the dataset, i.e., a test set (20 per cent), was used. The criteria to assess the predictive  
379 performance and reliability were set using *MSE*, *RMSE*, *MAE* and *R*<sup>2</sup>.

380 The three-layer feed-forward backpropagation ANN model provided the most  
381 satisfactory results compared to other regression models. The model yielded *MSE*,  
382 *RMSE* and *R*<sup>2</sup> values of 0.89, 0.93 and 0.82 in terms of 10-CV; 0.46, 0.67 and 0.90 for  
383 the validation set; and 0.47, 0.68 and 0.94 during the external validation using test set  
384 (Fig. 3, Table 5). The Levenberg-Marquardt (LM) algorithm used to build this model  
385 iteratively adjusts the model parameters to minimize the residual sum of squares  
386 between the model predictions and the observed data. At each iteration, the algorithm  
387 calculates the gradient and Hessian matrix of the objective function (which is the  
388 residual sum of squares) and then adjusts the model parameters by solving a modified  
389 system of equations that combines the Gauss-Newton method with the steepest  
390 descent method (Bilski et al., 2020). This technique hence results in the overall  
391 improvement of the model's generalisability.

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

393  $pLC_{50} = 3.25598 - 1.17895 B07.C.C.=0 + 3.97206e-14 B07.C.C.=1 - 0.03476 SMR -$   
 394  $0.660787 H.049 + 0.409287 MLFER\_S + 17.1359 VE1\_Dt - 262.482 VE2\_Dt -$   
 395  $0.0056275 ALOGP + 0.411825 SRW10 - 0.104173 C.002 + 0.596742 XLogP$

396 The LR model yielded satisfactory results for the 10-CV, with *MSE*, *RMSE*,  
 397 *MAE* and *R*<sup>2</sup> values of 1.48, 1.22, 0.94 and 0.69, respectively (Fig. 4a) and performed  
 398 better during the external validation with *MSE*, *RMSE*, *MAE* and *R*<sup>2</sup> value of 0.70, 0.84,  
 399 0.66 and 0.75, respectively (Fig. 4b). The good predictive performance of the LR  
 400 model could be due to employing Lasso regression technique, which adds  
 401 regularisation terms to the cost function to prevent overfitting and improve the  
 402 generalisability of the model (Yazdi et al., 2021).

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

412 On the other hand, the SVM model displayed slight overfitting on the training  
 413 set and underperformed compared to the other linear and non-linear regression  
 414 models yet produced moderate results. The model yielded *MSE*, *RMSE* and *R*<sup>2</sup> values  
 415 of 1.56, 1.25, 0.96 and 0.67, respectively, for the 10-CV (Fig. 4e) and 1.08, 1.04, 0.81  
 416 and 0.61 during the external validation (Fig. 4f). The possible explanation is, SVM

417 models are particularly susceptible to overfitting when the model has too many  
418 features relative to the size of the training data, leading to a sparse and high-  
419 dimensional feature space (Han & Jiang, 2014). Another reason could be that model's  
420 parameters, such as the regularisation parameter and the kernel function, are not  
421 chosen correctly (Han & Jiang, 2014).

422 Further, the summary and experimented pLC<sub>50</sub> versus predicted pLC<sub>50</sub>  
423 scatterplots are illustrated in Table 6 and Fig. 4 (a-f). An observation made on the  
424 measured and predicted biocide toxicity variation pattern in both training and validation  
425 sets suggests that all models performed reasonably well.

### 426 3.5. *Classification modelling*

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

439 The optimal DT model had the maximum number of splits as 100, 20 and 4,  
440 respectively, while the Gini's diversity index was employed as the split criterion. Each

441 model had the *ACC*, *SE* and *SP* value of 100% and *AUC* value of 1 for the 5-CV and  
442 test set, and as evident, performed the best for the classification of the three classes  
443 with no miscalculations. DT models, being non-parametric, do not make any  
444 assumptions about the distribution of the data. This makes them more flexible than  
445 parametric models like logistic regression, which assumes a linear relationship  
446 between the input features and the output (Abdalati et al., 2022).

447 In the case of optimal naïve Bayes, the model coupled with the Gaussian kernel  
448 performed reasonably well for the training set and performed better during the external  
449 validation. The model had the average *ACC*, *SE*, *SP* and *AUC* values of 91.5%, 75.8%,  
450 96.4% and 0.95, respectively; for 5-CV; and 94.4%, 97.8%, 96% and 0.94,  
451 respectively, for the test set. During the 5-fold cross-validation process, the naïve  
452 Bayes model was able to classify highly toxic biocides with 100% accuracy and no  
453 miscalculations, while 95% accuracy during the classification of moderately toxic  
454 compounds with three miscalculations and 91.5% accuracy during the classification of  
455 slightly/non-toxic compounds with three miscalculations. While during the external  
456 validation, the naïve Bayes model showed no miscalculations for the classification of  
457 moderately toxic and slightly/non-toxic biocides and only one miscalculation for the  
458 classification of highly toxic biocides. Naïve Bayes is, in general, a better classifier for  
459 similar tasks as it is robust to noise and irrelevant features because it assumes that  
460 features are independent of each other. This means that even if some features are not  
461 relevant to the classification task or contain noise, the classifier can still perform well  
462 (Salmi & Rustam, 2019).

463 However, it is essential to note that the overall generalisability and reliability of such  
464 classifiers in the regulatory context rely on the predictive performance with  
465 comparatively large and balanced datasets, which was a limiting factor in this study.

466 When evaluating the predictive performance of such models, it is also crucial to use  
467 appropriate metrics that accurately reflect the model's ability to predict the properties  
468 or activities of chemicals. Sensitivity, specificity, accuracy, and AUC can be less  
469 sensitive to class imbalance, but their performance can be affected by a class  
470 imbalance to some extent.

### 471 3.6. *Applicability domain (AD) assessment*

472 For reliable predictions, the applicability domain of the QSAR models was further  
473 analysed using the software Applicability Domain v1.0 which follows the  
474 standardization approach to probe any outliers present in training and test set.  
475 According to this method, if the standardised value of a compound's molecular  
476 descriptors is  $\leq 3$ , the compound is not an X-outlier or within AD, and *vice versa*. Only  
477 one compound in the test set was found to have an  $S_{new}$  value of 4.78, i.e.,  $> 3$   
478 (formaldehyde), suggesting an X-outlier or outside AD. While in the training set, four  
479 compounds had an  $S_{new}$  value of 3.15, 3.14, 5.33 and 3.28 (actane, dbnpa, neostanox  
480 and flubendiamide), implying X-outlier or outside the AD (appendix) (see  
481 Supplementary file 2). The outliers, nevertheless, were still incorporated during the  
482 model-building process due to fewer chemicals in the dataset, and the predictions  
483 were performed poorly for formaldehyde and neostanox only. This can be justified as  
484 only formaldehyde and neostanox had a considerably high  $S_{new}$  value, 4.78 and 5.33,  
485 respectively. A possible explanation for the detection of formaldehyde as an outlier in  
486 the test set is its relatively simple structure in comparison to the majority with highly  
487 diverse and complex structures. In addition, formaldehyde also had the lowest atomic  
488 LogP value (ALOGP), suggesting higher hydrophilicity and one hydrogen atom (H-  
489 049) directly attached to the carbon atom (C1) in formaldehyde, while one hydrogen

490 atom (H-049) attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp) of another molecule. In the  
491 training set, neostanox had exceedingly high atomic LogP, suggesting a very high  
492 hydrophobic nature; this is due to the presence of non-polar functional groups, also  
493 resulting in high Atom-Type E-state (ATE). The relationship between ATE and logP is  
494 based on the fact that the electronic state of atoms in a molecule can influence the  
495 molecule's solubility and partitioning behaviour. In particular, atoms with higher ATE  
496 values (indicating a more electron-withdrawing or polar group) tend to be more  
497 hydrophilic and less likely to partition into non-polar solvents (Kier et al., 1999). In  
498 addition, neostanox was the only chemical with the presence of an [Sn] atom in the  
499 dataset. The presence of [Sn] molecular descriptor in the case of neostanox can  
500 significantly distinguish the substance from the dataset, eventually affecting the overall  
501 generalisability of the silico models. The other possible reason for the poor predictive  
502 performance of molecularly similar compounds could be factors such as erroneous,  
503 insufficient or poor-quality raw data used for training the model. Hence, it is  
504 recommended to exclude the detected outliers from the dataset in order to improve  
505 the overall generalisability and predictive performance of the model.

### 506 *3.7. Adaptive modelling for reliable ecotoxicological evaluations in a* 507 *regulatory context*

508 The developed ML models presented in this report have shown good predictive  
509 performance, high generalisability, and the potential to replace animal testing for  
510 biocide ecotoxicological screening in marine crustaceans. However, its acceptance  
511 and the impact it merits in regulatory decision-making is still a topic of debate. The key  
512 arguments are (i) model generalisability and adaptability (ii) reliability of model  
513 validation (iii) confidence in predictive accuracy and (iv) transparency and



514 interpretability of some ML algorithms. The OECD guidelines principle 2 provides  
515 important guidance on the quality and relevance of data used in chemical safety  
516 assessments. However, there are some limitations to its implementation, such as the  
517 limited availability of high-quality (LC<sub>50</sub>) datasets for many chemicals. In some cases,  
518 there may be gaps in the data, or the available data may not be sufficient to fully  
519 characterize the risks associated with a chemical.

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

534 Experimental validation is also an essential step in the development and evaluation  
535 of QSAR models for regulatory purposes. This validation process involves testing the  
536 model's predictions against experimental data to evaluate its accuracy and reliability  
537 (OECD, 2004). While experimental validation is certainly an important part of validating  
538 any scientific model or theory, it is not always feasible or necessary for QSAR models

539 (Tropsha, 2010). This is because QSAR models are based on statistical relationships  
540 between chemical structures and biological activities. These relationships can be  
541 tested using various statistical measures, such as sensitivity, specificity, accuracy,  
542 precision, and the area under the receiver operating characteristic (ROC) curve  
543 (Grandini et al., 2020). These metrics provide information on the models' ability to  
544 correctly predict positive and negative cases and to distinguish between hazardous  
545 and non-hazardous chemicals. In addition, experimental validation can be time-  
546 consuming, costly, and sometimes unethical if it involves animal testing. QSAR models  
547 offer a faster, cheaper, and more ethical alternative to experimental testing. They can  
548 also be used to prioritise chemicals for further testing or to design new chemicals with  
549 specific properties, which can help to reduce the need for animal testing (Khan et al.,  
550 2019).

551 In our study, we employed k-fold cross-validation, where the entire dataset was  
552 divided into ten subsets, of which nine subset was used to train the model and the  
553 remaining subset was treated as a test set to validate the model. This method  
554 improves the robustness of the model to data variability by averaging the performance  
555 across multiple runs of the cross-validation process. This can help to reduce the  
556 impact of data variability on the model's predictive performance. A similar approach  
557 was adopted by Liu et al. (2019) to predict and validate chemical toxicity in marine  
558 crustaceans, where the classification models yielded fairly well results. Furthermore,  
559 for multi-class classification modelling, where the dataset is relatively small, and one  
560 class is more prevalent. It is important to use a combination of evaluation metrics,  
561 including those less sensitive to class imbalance. For example, Singh et al. (2013)  
562 employed a combination of sensitivity, specificity and accuracy, which measures the  
563 occurrence of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false

564 negatives (*FN*) in the multi-class classification of diverse chemicals acute toxicity in  
565 fish. A similar approach was also adopted by Liu et al. (2019) to classify acute chemical  
566 toxicity in marine crustaceans. Various other multi-class classification evaluation  
567 metrics such as Matthews Correlation Coefficient (*MCC*), Cohen's Kappa, macro-  
568 averaged precision, recall, and F1-score can also provide a more accurate  
569 assessment of the model's predictive performance in the presence of class imbalance  
570 (Grandini et al., 2020).

### 571 3.8. Comparison of developed models with models available in the 572 literature

573 The  $LC_{50}$  is a widely used endpoint in QSAR modelling, particularly in the field of  
574 ecotoxicology. Such QSAR models that predict  $LC_{50}$  values can provide valuable  
575 information for regulatory decision-making and environmental risk assessment (ERA).  
576 However, the literature survey showed that the potential of computational models to  
577 predict biocide  $LC_{50}$  in marine crustaceans had not yet been extensively explored.  
578 Therefore, a quantitative comparison with others' work would be irrelevant because  
579 the datasets and target organisms differ between the models. Nonetheless, a simple  
580 comparison of our model methodology and result statistics will give fundamental  
581 insight into the accuracy of various approaches to building such models.

582 Various classification-based models were developed by Liu et al. (2019) to predict  
583 and classify the  $LC_{50}$  values of a wide array of chemicals in marine crustaceans. The  
584 method employed six ML models, which are SVM, NB, RF, DT, kNN, and ANN, and  
585 trained using a set of 1D/2D molecular descriptors and fingerprints. Similar 10-fold  
586 cross-validation was also employed for model validation, and the *AUC* values of the  
587 developed models ranged from 0.80 – 0.90 for test sets. The DT model developed in

588 our study showed the *AUC* value of 1 for both the training and test set. However, It is  
589 important to note that the models developed by Liu et al. (2019) used a significantly  
590 large dataset (>1000) which was a limiting factor in our study. For the acceptance of  
591 a model in a regulatory context, it is also recommended that the models are trained  
592 using a large and good-quality dataset. Similarly, two partial least squares (PLS)  
593 regression-based models were developed by Khan et al. (2019) to predict LC<sub>50</sub> values  
594 of biocides in *Daphnia magna* and fish toxicities using 2D descriptors. The method  
595 employed leave-one-out cross-validation to validate the models, and the results  
596 yielded *R*<sup>2</sup> of 0.80 and 0.64, respectively, for fish training and test set, and *R*<sup>2</sup> 0.87 and  
597 0.81, respectively, for *Daphnia magna* training and test set. These models showed  
598 satisfactory results; however, they tend to overfit the training set. Overfitting occurs  
599 when a model learns the patterns in the training data too well and becomes too specific  
600 to that data. As a result, the model may not generalize well to new, unseen data, such  
601 as the test set. The presented models in our study have shown high generalisability  
602 by avoiding overfitting on the training data suggesting appropriateness to replace  
603 unnecessary animal testing to predict biocide toxicity in a wide range of marine  
604 crustacean species.

#### 605 **4. Conclusions**

606 In this study, firstly, an overview was presented on how extensive use of biocidal  
607 products can have a detrimental impact on the aquatic organisms, with particular  
608 reference to crustaceans due to their non-target mechanism of action. Secondly, in  
609 the light of incorporating animal alternatives for environmental risk assessment (ERA)  
610 of hazardous chemicals, *in silico* models were built to fill this data gap by predicting  
611 the acute chemical toxicity of biocidal chemicals in environmentally sensitive

612 invertebrates - marine crustaceans. The work presented herein has shown that *in silico*  
613 modelling approaches are a powerful method to predict acute chemical toxicity of  
614 biocides, enabling rapid prioritisation of compounds during ERA. The biocide dataset  
615 used in the research shows good diversity, and each predictive model is quite diverse  
616 in its approach, as well. All six models in this study yielded satisfactory results, and  
617 the feed-forward backpropagation-based artificial neural network model showed the  
618 best performance during regression analysis, while decision tree model performed the  
619 best for the classification of different toxicities. Nevertheless, ML approaches have  
620 great potential in ecotoxicological studies, and further improvement and understanding  
621 of the underlying science are important. The major limiting factor in this study to build  
622 an even more robust model was the small biocide sample size of the dataset ( $n=89$ );  
623 hence, updating the chemical and ecotoxicological databases is also pivotal. In  
624 addition to predicting the toxicity of a particular chemical, ML can also be used to  
625 interpret the influence of a particular molecular descriptor or property contributing to  
626 its toxicity, allowing to manufacture of a greener and more sustainable chemical  
627 product. The developed models are capable of predicting the toxicities of untested  
628 biocides within the applicability domain of the models.

#### 629 **Declaration of competing interest:**

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

#### 632 **Acknowledgement**

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

636 **Krishnan et al.**

637 **Author contributions**

638

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

641 **IH:** Methodology, Software, Validation, Supervision, Review and editing.

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

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

645 **Reference**

646 Abdalati, A., Saed, A., & Jaharadak, A. A. (2022). Implementation with Performance

647 Evaluation of Decision Tree Classifier for Uncertain Data: Literature Review.

648 *International Journal of Multidisciplinary Research and Publications (IJMRAP),*

649 5(5), 125–132.

650 Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model

651 selection. *Statistics Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>

652 Backhaus, T., Altenburger, R., Faust, M., Frein, D., Frische, T., Johansson, P.,

653 Kehrer, A., & Porsbring, T. (2013). Proposal for environmental mixture risk

654 assessment in the context of the biocidal product authorization in the EU.

655 *Environmental Sciences Europe*, 25(1), 1–9. <https://doi.org/10.1186/2190-4715->

656 25-4/FIGURES/2

657 Barros, R. P. C., Sousa, N. F., Scotti, L., & Scotti, M. T. (2020). Use of machine

658 learning and classical QSAR methods in computational ecotoxicology. *Methods*

659 *in Pharmacology and Toxicology*, 151–175. <https://doi.org/10.1007/978-1-0716->

660 0150-1\_7

661 Berthold, M. R., Cebron, N., Dill, F., Di Fatta, G., Gabriel, T. R., Georg, F., Meinl, T.,

662 Ohl, P., Sieb, C., & Wiswedel, B. (2009). KNIME - the Konstanz information

663 miner. *ACM SIGKDD Explorations Newsletter*, 58–61.

664 <https://doi.org/10.1145/1656274.1656280>

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

668 Chandrasekaran, B., Abed, S. N., Al-Attraqchi, O., Kuche, K., & Tekade, R. K.  
669 (2018). Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties.  
670 *Dosage Form Design Parameters*, 2, 731–755. <https://doi.org/10.1016/B978-0->  
671 12-814421-3.00021-X

672 Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.  
673 *Proceedings of the ACM SIGKDD International Conference on Knowledge*  
674 *Discovery and Data Mining, 13-17-August-2016*, 785–794.  
675 <https://doi.org/10.1145/2939672.2939785>

676 COMBASE. (2016). *COMBASE*. <https://www.life-combase.com/index.php/en/>

677 Coors, A., Vollmar, P., Heim, J., Sacher, F., & Kehrer, A. (2018). Environmental risk  
678 assessment of biocidal products: identification of relevant components and  
679 reliability of a component-based mixture assessment. *Environmental Sciences*  
680 *Europe*, 30(1), 1–15. <https://doi.org/10.1186/S12302-017-0130-0/TABLES/4>

681 Dale, V. H., & Beyeler, S. C. (2001). Challenges in the development and use of  
682 ecological indicators. *Ecological Indicators*, 1(1), 3–10.  
683 [https://doi.org/10.1016/S1470-160X\(01\)00003-6](https://doi.org/10.1016/S1470-160X(01)00003-6)

684 Demšar, J., Erjavec, A., Hočevár, T., Milutinovič, M., Možina, M., Toplak, M., Umek,  
685 L., Zbontar, J., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python

686 Tomaz Curk Matija Polajnar Lañ Zagar. *Journal of Machine Learning Research*,  
687 14, 2349–2353.

688 Devi, T. G., Patil, N., Rai, S., & Sarah, C. P. (2023). Segmentation and classification  
689 of white blood cancer cells from bone marrow microscopic images using duplet-  
690 convolutional neural network design. *Multimedia Tools and Applications*, 1–23.  
691 <https://doi.org/10.1007/S11042-023-14899-9/METRICS>

692 Ebenuwa, S. H., Sharif, M. S., Alazab, M., & Al-Nemrat, A. (2019). Variance Ranking  
693 Attributes Selection Techniques for Binary Classification Problem in Imbalance  
694 Data. *IEEE Access*, 7, 24649–24666.  
695 <https://doi.org/10.1109/ACCESS.2019.2899578>

696 EC. (2009). Assessment of different options to address risks from the use phase of  
697 biocides *Final report*. [www.cowi.com](http://www.cowi.com)

698 EC. (2018). Report from the commission to the European parliament and the council  
699 on the implementation of the Union authorisation of biocidal products in  
700 accordance with Article 42(3) of Regulation (EU) No 528/2012 of the European  
701 Parliament and of the Council concerning the making available on the market  
702 and use of biocidal products. [https://ec.europa.eu/health/biocides/regulation\\_en](https://ec.europa.eu/health/biocides/regulation_en)

703 ECHA. (2022). Homepage - *ECHA*. <https://echa.europa.eu/>

704 EU. (2012). Regulation (EU) No 528/2012 of the European Parliament and of the  
705 Council of 22 May 2012 Concerning the Making Available on the Market and  
706 use of Biocidal Products. *ISSN 1977 677, 2985*.

707 Flemming, H.-C., Murthy, P. S., Venkatesan, R., & Cooksey, K. (2009). *Marine and*  
708 *industrial biofouling* (Vol. 333). Springer.



709 Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: an  
710 Overview. <https://arxiv.org/abs/2008.05756v1>

711 Han, H., & Jiang, X. (2014). Overcome support vector machine diagnosis overfitting.  
712 *Cancer Informatics*, 13, CIN-S13875.

713 Hansch, C., & Fujita, T. (1964).  $\rho$ - $\sigma$ - $\pi$  Analysis. A Method for the Correlation of  
714 Biological Activity and Chemical Structure. *Journal of the American Chemical*  
715 *Society*, 86(8), 1616–1626.  
716 [https://doi.org/10.1021/JA01062A035/ASSET/JA01062A035.FP.PNG\\_V03](https://doi.org/10.1021/JA01062A035/ASSET/JA01062A035.FP.PNG_V03)

717 Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S. A., Attene-Ramos, M.,  
718 Zhao, T., Austin, C. P., & Simeonov, A. (2016). Modelling the Tox21 10 K  
719 chemical profiles for toxicity prediction and mechanism characterization. *Nature*  
720 *Communications* 2016 7:1, 7(1), 1–10. <https://doi.org/10.1038/ncomms10425>

721 Isabona, J., Imoize, A. L., & Kim, Y. (2022). Machine Learning-based boosted  
722 regression ensemble combined with hyperparameter tuning for optimal adaptive  
723 learning. *Sensors*. 22, 3776, 22(10), 3776. <https://doi.org/10.3390/S22103776>

724 Kensert, A., Alvarsson, J., Norinder, U., & Spjuth, O. (2018). Evaluating parameters  
725 for ligand-based modeling with random forest on sparse data sets. *Journal of*  
726 *Cheminformatics*, 10(1), 49. <https://doi.org/10.1186/S13321-018-0304-9>

727 Khan, K., Khan, P. M., Lavado, G., Valsecchi, C., Pasqualini, J., Baderna, D., Marzo,  
728 M., Lombardo, A., Roy, K., & Benfenati, E. (2019). QSAR modeling of *Daphnia*  
729 *magna* and fish toxicities of biocides using 2D descriptors. *Chemosphere*, 229,  
730 8–17. <https://doi.org/10.1016/J.CHEMOSPHERE.2019.04.204>

731 Kier, L. B., Hall, L. H., & 1937-. (1999). *Molecular structure description*. 41.  
732 <https://doi.org/10.3/JQUERY-UI.JS>

733 Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A.,  
734 Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2021).  
735 PubChem in 2021: new data content and improved web interfaces. *Nucleic*  
736 *Acids Research*, 49(D1), D1388–D1395. <https://doi.org/10.1093/NAR/GKAA971>

737 Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal*  
738 *of Statistical Software*, 28(5), 1–26. <https://doi.org/10.18637/JSS.V028.I05>

739 Langdon, C. J., Vance, M. M., Harmon, V. L., Kreeger, K. E., Kreeger, D. A., &  
740 Chapman, G. A. (1996). A 7-D toxicity test for marine pollutants using the pacific  
741 mysid *Mysidopsis intii*. 1. Culture and protocol development. *Environmental*  
742 *Toxicology and Chemistry*, 15(10), 1815–1823.  
743 <https://doi.org/10.1002/ETC.5620151024>

744 Liu, L., Yang, H., Cai, Y., Cao, Q., Sun, L., Wang, Z., Li, W., Liu, G., Lee, P. W., &  
745 Tang, Y. (2019). *In silico* prediction of chemical aquatic toxicity for marine  
746 crustaceans via machine learning. *Toxicology Research*, 8(3), 341–352.  
747 <https://doi.org/10.1039/c8tx00331a>

748 Liu, R., Madore, M., Glover, K. P., Feasel, M. G., & Wallqvist, A. (2018). Assessing  
749 deep and shallow learning methods for quantitative prediction of acute chemical  
750 toxicity. *Toxicological Sciences*, 164(2), 512–526.  
751 <https://doi.org/10.1093/toxsci/kfy111>

752 Lussier, S. M., Kuhn, A., & Comeleo, R. (1999). An evaluation of the seven-day  
753 toxicity test with *Americamysis bahia* (formerly *Mysidopsis bahia*).  
754 *Environmental Toxicology and Chemistry*, 18(12), 2888–2893.  
755 <https://doi.org/10.1002/ETC.5620181233>

756 Marzo, M., Lavado, G. J., Como, F., Toropova, A. P., Toropov, A. A., Baderna, D.,  
757 Cappelli, C., Lombardo, A., Toma, C., Blázquez, M., & Benfenati, E. (2020).  
758 QSAR models for biocides: The example of the prediction of *Daphnia magna*  
759 acute toxicity. *SAR and QSAR in Environmental Research*, 31(3), 227–243.  
760 [https://doi.org/10.1080/1062936X.2019.1709221/SUPPL\\_FILE/GSAR\\_A\\_17092](https://doi.org/10.1080/1062936X.2019.1709221/SUPPL_FILE/GSAR_A_17092)  
761 [21\\_SM4833.DOCX](https://doi.org/10.1080/1062936X.2019.1709221/SUPPL_FILE/GSAR_A_17092)

762 MATLAB. (2010). *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks  
763 Inc.

764 Mauri, A., & Srl, A. (2021). Development of software tools for the application of  
765 QSAR models View project OpenTox View project Chapter 32 alvaDesc: A Tool  
766 to Calculate and Analyze Molecular Descriptors and Fingerprints.  
767 [https://doi.org/10.1007/978-1-0716-0150-1\\_32](https://doi.org/10.1007/978-1-0716-0150-1_32)

768 Miller, T. H., Gallidabino, M. D., Macrae, J. I., Hogstrand, C., Bury, N. R., Barron, L.  
769 P., Snape, J. R., & Owen, S. F. (2018). Machine Learning for Environmental  
770 Toxicology: A Call for Integration and Innovation. In *Environmental Science and*  
771 *Technology* (Vol. 52, Issue 22, pp. 12953–12955). American Chemical Society.  
772 <https://doi.org/10.1021/acs.est.8b05382>

773 Miller, T. H., Gallidabino, M. D., MacRae, J. R., Owen, S. F., Bury, N. R., & Barron,  
774 L. P. (2019). Prediction of bioconcentration factors in fish and invertebrates  
775 using machine learning. *Science of the Total Environment*, 648, 80–89.  
776 <https://doi.org/10.1016/j.scitotenv.2018.08.122>

777 Oberdörster, E., & Cheek, A. O. (2001). Gender benders at the beach: Endocrine  
778 disruption in marine and estuarine organisms. *Environmental Toxicology and*  
779 *Chemistry*, 20(1), 23–36. <https://doi.org/10.1002/ETC.5620200103>

780 OECD. (2004). *Validation of (Q)SAR Models*.  
781 <https://www.oecd.org/chemicalsafety/risk->  
782 [assessment/validationofqsarmodels.htm](https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm)

783 Olker, J. H., Elonen, C. M., Pilli, A., Anderson, A., Kinziger, B., Erickson, S.,  
784 Skopinski, M., Pomplun, A., LaLone, C. A., Russom, C. L., & Hoff, D. (2022).  
785 The ECOTOXicology Knowledgebase: A Curated Database of Ecologically  
786 Relevant Toxicity Tests to Support Environmental Research and Risk  
787 Assessment. *Environmental Toxicology and Chemistry*, 41(6), 1520–1539.  
788 <https://doi.org/10.1002/ETC.5324>

789 Rand, G. M. (1985). Introduction. IN: Rand, GM, Petrocelli, SR Fundamentals of  
790 aquatic toxicology: methods and application. *London, Hemisphere Publishing*  
791 *Corporation. Cap, 1, 1–28.*

792 Roast, S. D., Thompson, R. S., Donkin, P., Widdows, J., & Jones, M. B. (1999).  
793 Toxicity of the organophosphate pesticides chlorpyrifos and dimethoate to  
794 *Neomysis integer* (Crustacea: Mysidacea). *Water Research*, 33(2), 319–326.  
795 [https://doi.org/10.1016/S0043-1354\(98\)00248-6](https://doi.org/10.1016/S0043-1354(98)00248-6)

796 Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of*  
797 *Chemical Information and Modelling*, 50(5), 742–754.  
798 <https://doi.org/10.1021/CI100050T/ASSET/IMAGES/MEDIUM/CI-2010->  
799 [00050T\\_0018.GIF](https://doi.org/10.1021/CI100050T/ASSET/IMAGES/MEDIUM/CI-2010-00050T_0018.GIF)

800 Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining  
801 applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory*  
802 *Systems*, 145, 22–29. <https://doi.org/10.1016/J.CHEMOLAB.2015.04.013>

803 Russom, C. L., Bradbury, S. P., Broderius, S. J., Hammermeister, D. E., &  
804 Drummond, R. A. (1997). Predicting modes of toxic action from chemical  
805 structure: Acute toxicity in the fathead minnow (*Pimephales promelas*).  
806 *Environmental Toxicology and Chemistry*, 16(5), 948–967.  
807 <https://doi.org/10.1002/ETC.5620160514>

808 Salmi, N., & Rustam, Z. (2019). Naïve Bayes Classifier Models for Predicting the  
809 Colon Cancer. *IOP Conference Series: Materials Science and Engineering*,  
810 546(5), 052068. <https://doi.org/10.1088/1757-899X/546/5/052068>

811 Schüürmann, G., Ebert, R. U., & Kühne, R. (2011). Quantitative read-across for  
812 predicting the acute fish toxicity of organic compounds. *Environmental Science*  
813 *and Technology*, 45(10), 4616–4622. <https://doi.org/10.1021/ES200361R>

814 Sieg, J., Flachsenberg, F., & Rarey, M. (2019). In Need of Bias Control: Evaluating  
815 Chemical Data for Machine Learning in Structure-Based Virtual Screening.  
816 *Journal of Chemical Information and Modelling*, 59(3), 947–961.  
817 [https://doi.org/10.1021/ACS.JCIM.8B00712/SUPPL\\_FILE/CI8B00712\\_SI\\_001.PDF](https://doi.org/10.1021/ACS.JCIM.8B00712/SUPPL_FILE/CI8B00712_SI_001.PDF)

818 DF

819 Singh, K. P., Gupta, S., & Rai, P. (2013). Predicting acute aquatic toxicity of  
820 structurally diverse chemicals in fish using artificial intelligence approaches.  
821 *Ecotoxicology and Environmental Safety*, 95, 221–233.  
822 <https://doi.org/10.1016/j.ecoenv.2013.05.017>

823 Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W.,  
824 Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., Todeschini,  
825 R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J.,  
826 Schwab, C., Baskin, I. I., ... Tetko, I. V. (2011). Online chemical modeling

827 environment (OCHEM): web platform for data storage, model development and  
828 publishing of chemical information. *Journal of Computer-Aided Molecular*  
829 *Design*, 25(6), 533–554. <https://doi.org/10.1007/S10822-011-9440-2>

830 Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and  
831 Exploitation. *Molecular Informatics*, 29(6–7), 476–488.  
832 <https://doi.org/10.1002/MINF.201000061>

833 US-EPA. (2021). *Technical Overview of Ecological Risk Assessment - Analysis*  
834 *Phase: Ecological Effects Characterization | US EPA*.  
835 [https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/technical-](https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/technical-overview-ecological-risk-assessment-0)  
836 [overview-ecological-risk-assessment-0](https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/technical-overview-ecological-risk-assessment-0)

837 Walters, P. (2019, November 1). *Visualizing Chemical Space*.  
838 [http://practicalcheminformatics.blogspot.com/2019/11/visualizing-chemical-](http://practicalcheminformatics.blogspot.com/2019/11/visualizing-chemical-space.html)  
839 [space.html](http://practicalcheminformatics.blogspot.com/2019/11/visualizing-chemical-space.html)

840 Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., & Yu, Y.  
841 (2018). *Taking Human out of Learning Applications: A Survey on Automated*  
842 *Machine Learning*. <http://arxiv.org/abs/1810.13306>

843 Yazdi, M., Golilarz, N. A., Nedjati, A., & Adesina, K. A. (2021). An improved lasso  
844 regression model for evaluating the efficiency of intervention actions in a system  
845 reliability analysis. *Neural Computing and Applications* 2021 33:13, 33(13),  
846 7913–7928. <https://doi.org/10.1007/S00521-020-05537-8>

847

848

849 **Figure legends:**

850

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

853 **Figure 2:** Figures illustrating diversity in the dataset: (a) ALOGP molecular descriptor  
854 correlation with experimental toxicity pLC<sub>50</sub> mmol/l. (b) Chemical space of biocide  
855 dataset defined using principal component analysis (PCA). The colours and sizes  
856 represent the varying pLC<sub>50</sub> mmol/l values of biocides in the dataset. (c) Frequency and  
857 distribution of biocides (blue bar) in the marine crustacean toxicity dataset according  
858 to their toxicity (pLC<sub>50</sub> mmol/l). (d) Tanimoto similarity index heatmap of the biocidal  
859 compounds in the dataset using 2D circular Morgan fingerprints. The similarity index  
860 increases from zero to one.

861 **Figure 3:** Scatterplot of the experimented and model predicted values of biocide  
862 toxicity (pLC<sub>50</sub>) in the training set, validation set, test set and complete set of ANN  
863 model.

864 **Figure 4:** Regression scatter plots for training and test sets of machine learning  
865 models (a-b) LR, (c-d) RF, (e-f) SVM, respectively, used in this study (Experimental  
866 pLC<sub>50</sub> – x-axis vs. Predicted pLC<sub>50</sub> – y-axis).

867

868 **Table Captions**

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

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

871 **Table 3:** Chemical toxicity categories in marine organisms.

872 **Table 4:** Molecular descriptors used for model building.

873 **Table 5:** Performance parameters for ANN regression model to predict acute toxicity  
874 of biocides.

875 **Table 6:** Performance parameters for various regression models to predict acute  
876 toxicity of biocides.

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

879

880 **Captions for Supplementary Materials**

881 S1. Biocide acute chemical toxicity in marine crustaceans dataset used in this study.

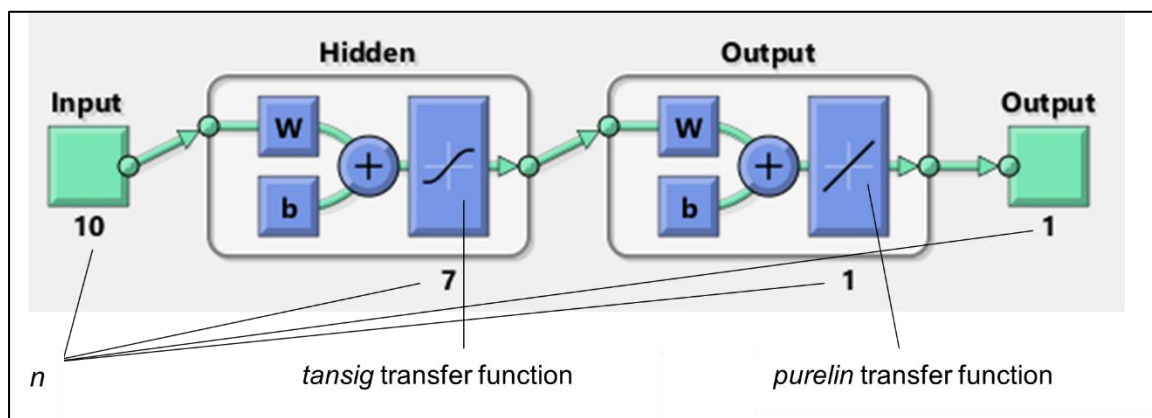
882 S2. Applicability Domain Training set.

883 S3. Molecular descriptors selected for regression analysis.

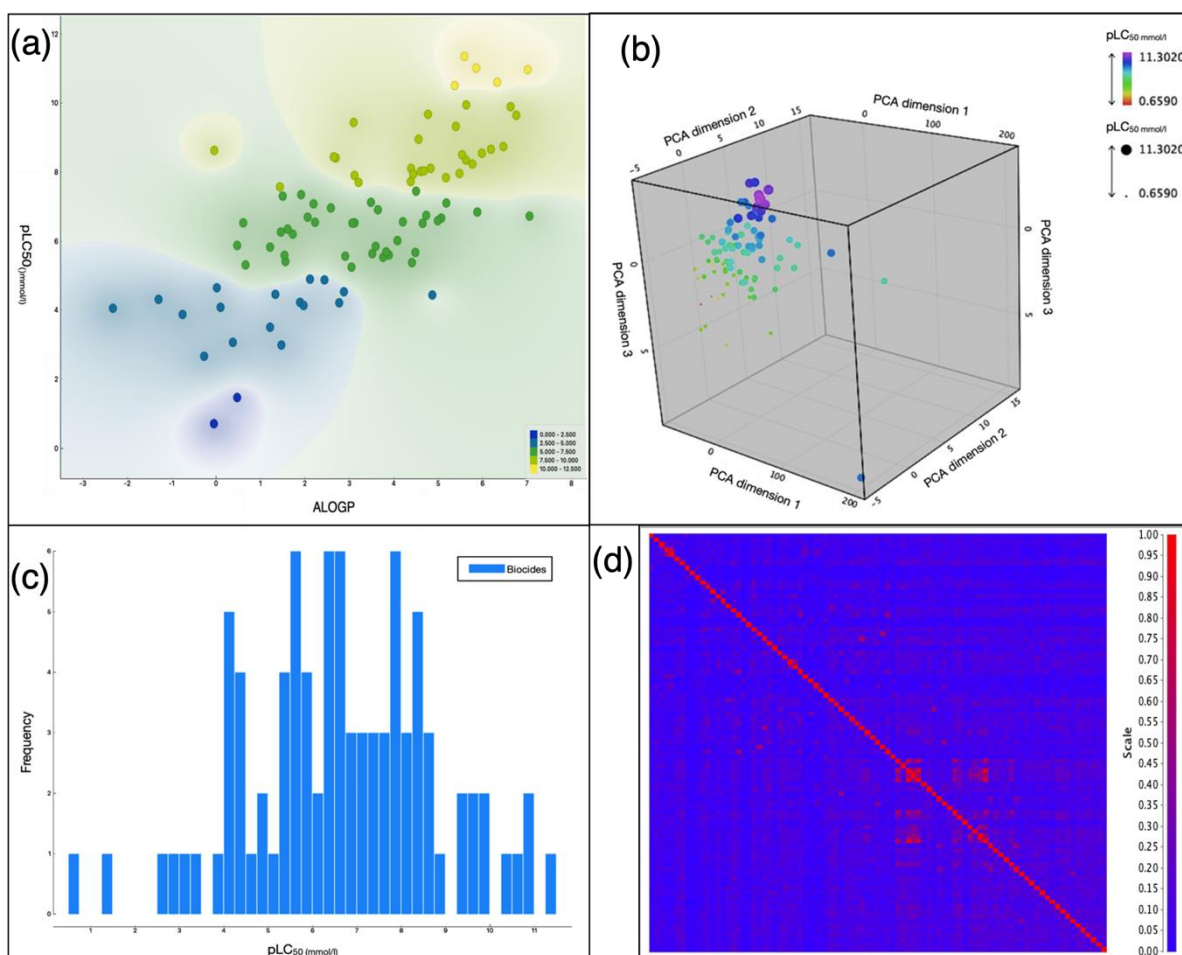
884 S4. Molecular descriptors selected for classification analysis.

885 S5. Best Subset Selection (BSS) for highest correlation features for regression  
886 analysis.

887



890 **Figure 2:** ANN architecture used for model building ( $n$  = number of neurons used in  
891 each layer,  $W$  = weight vector and  $b$  = bias).



893 **Figure 2:** Figures illustrating diversity in the dataset: (a) ALOGP molecular descriptor  
894 correlation with experimental toxicity  $pLC_{50}$  mmol/l. (b) Chemical space of biocide  
895 dataset defined using principal component analysis (PCA). The colours and sizes  
896 represent the varying  $pLC_{50}$  mmol/l values of biocides in the dataset. (c) Frequency and  
897 distribution of biocides (blue bar) in the marine crustacean toxicity dataset according



898 to their toxicity ( $pLC_{50}$  mmol/l). (d) Tanimoto similarity index heatmap of the biocidal  
899 compounds in the dataset using 2D circular Morgan fingerprints. The similarity index  
900 increases from zero to one.

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

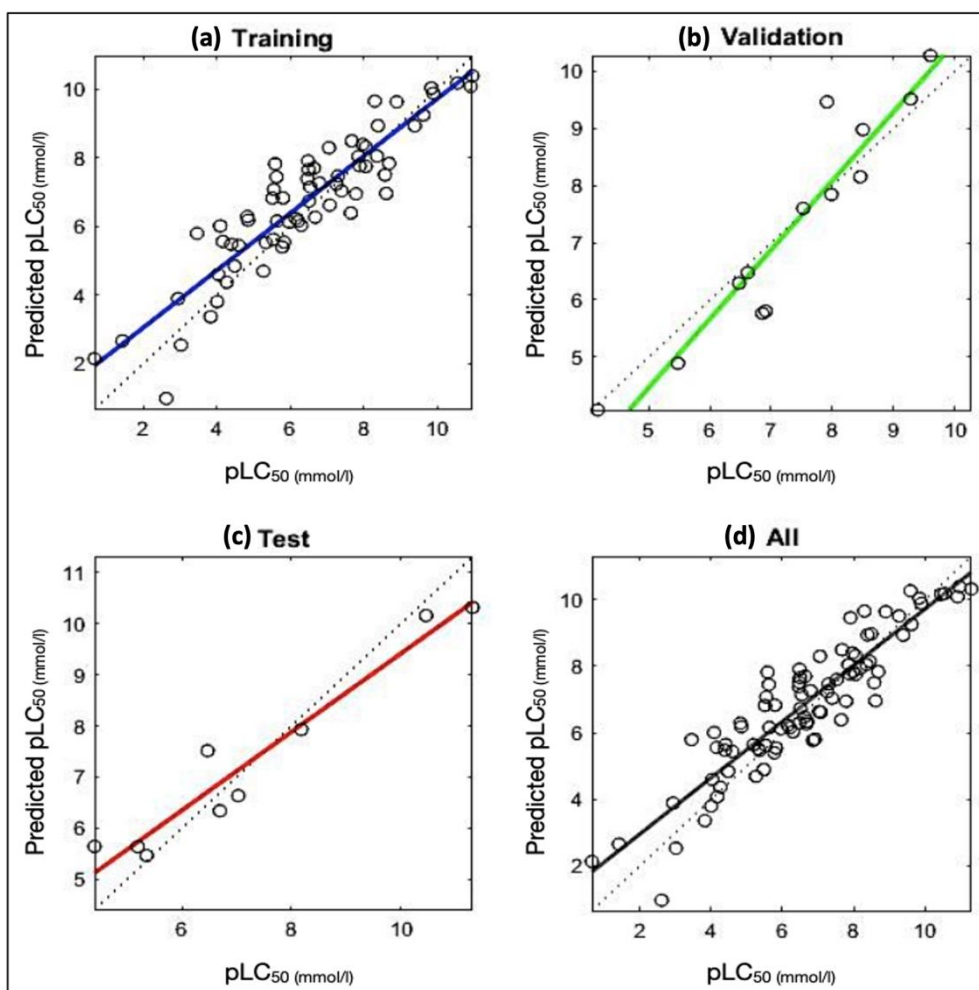
916

917

918

919

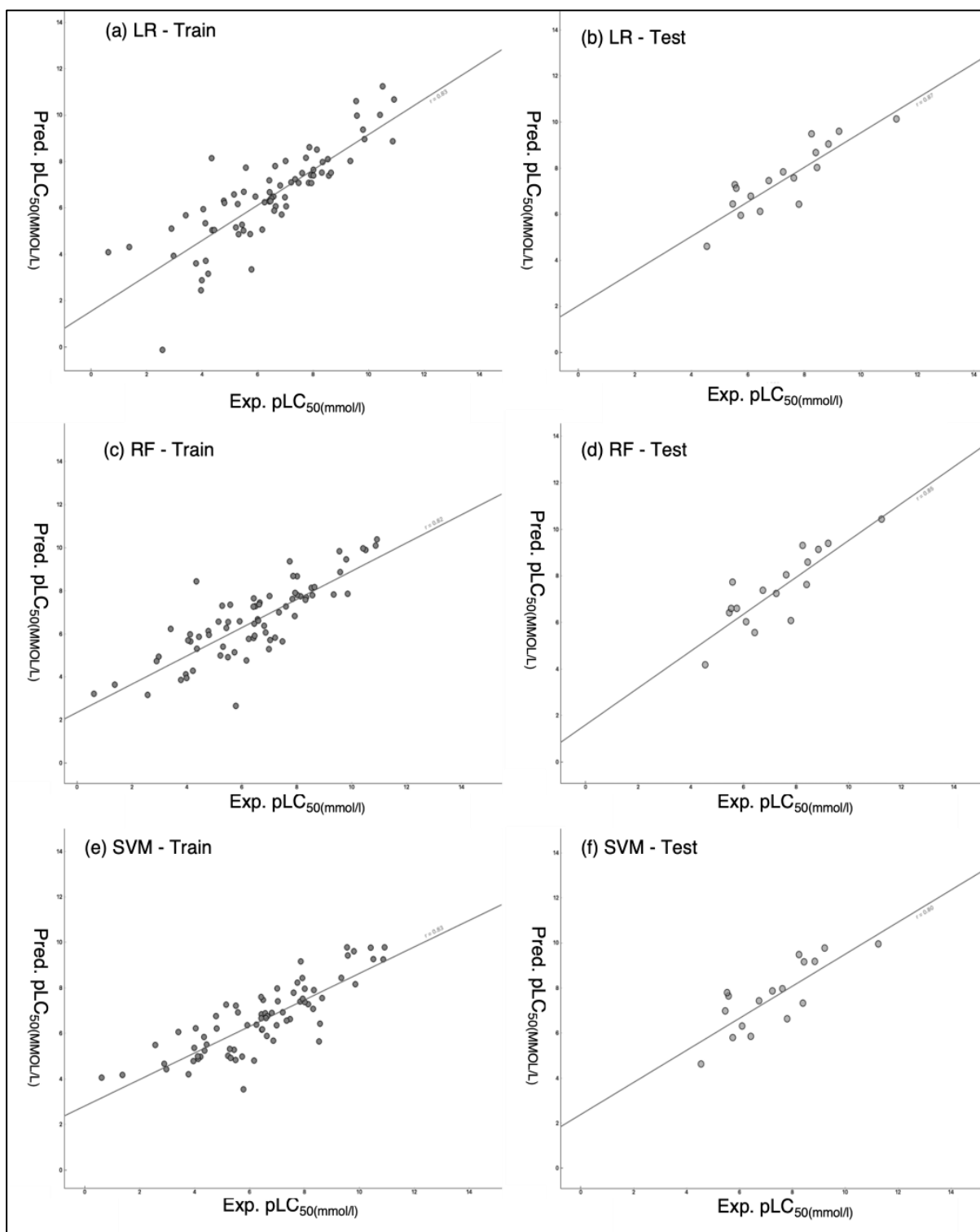
920



921 **Figure 3:** Scatterplot of the experimented and model predicted values of biocide  
922 toxicity ( $pLC_{50}$ ) in the (a) training set (b) validation set (c) test set and (d) complete  
923 set of ANN model.

924

925



926 **Figure 4:** Regression scatter plots for training and test sets of machine learning  
 927 models (a-b) LR, (c-d) RF, (e-f) SVM, respectively, used in this study (Experimental  
 928 pLC<sub>50</sub> – x-axis vs. Predicted pLC<sub>50</sub> – y-axis).

929  
 930

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

Analysis	Model	Equation	Hyperparameter	Referen
Regression	SVM	$K(X_1, X_2) = \exp\left(-\frac{\ X_1 - X_2\ ^2}{2\sigma^2}\right)$	•RBF Kernel	Chang al., 2016
	RF	$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x^l)$	•No. of trees: 10 •No. of attributes in each split: 8	Ho, 1999
	LR	$Y_i = f(X_i, \beta) + e_i$	•Lasso regression • $\alpha = 0.0001$	Cohen al., 2014
	ANN	$g(x) = f^L(W^L f^{L-1}(W^{L-1} \dots f^1(W^1 x) \dots))$	•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, 1999
	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 al., 2007

RBF – radial basis function,  $\sigma$ - variance, X1 and X2 – two points, K – kernel function, B – bagging,  $x^l$ - test samples,  $b = 1, f_b$ - trees,  $Y_i$ - dependent variable,  $f$ - function,  $X_i$ - independent variable,  $\beta$ - 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,  $\sigma$ - independent variable

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

933

Analysis	Statistical estimator	Theory	Equation	Reference
Regression	MSE	Average squared difference between predicted value and actual value	$MSE = \frac{1}{n} \sum_{i=0}^n (Y_i - \hat{Y}_i)^2$	Bickel et al., 2015.
	RMSE/ RMSD	Standard deviation of prediction errors	$RMSE = \sqrt{\frac{\sum_{i=1}^N (x_i - \hat{x}_i)^2}{N}}$	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 <sup>2</sup>	Variation in prediction proposed by the model	$R^2 = 1 - \frac{RSS}{TSS}$	Damodar, 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

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

938 **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	

939  
940  
941

**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	octanol/water partition coefficients of organic compounds	XLogP	
	MLFER_S	PaDEL 2	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	Ghose-Crippen LogKow	ALogP	
	ATSC1c	PaDEL 2	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		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	RDKit	MOE logP VSA Descriptor 10 (0.40 ≤ x < 0.50)	molecular surface area descriptors
smr_VSA3		MOE MR VSA Descriptor 3 (1.82 ≤ x < 2.24)	

942

943

944 **Table 5:** Performance parameters for ANN regression model to predict acute toxicity  
945 of biocides.

Model	Dataset	No. of compound	MSE	RMSE	R <sup>2</sup>
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

946

947 **Table 6:** Performance parameters for various regression models to predict acute  
948 toxicity of biocides.

Model	Dataset	MSE	RMSE	MAE	R <sup>2</sup>
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

949

950

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

Training set (5-fold Cross-Validation)								
Decision Tree	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
	<b>Total</b>	<b>71</b>						

Naïve Bayes

<b>Test set (external validation)</b>							
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
<b>Total</b>	18						

<b>Training set (5-fold Cross-Validation)</b>							
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
<b>Total</b>	71						

<b>Test set (external validation)</b>							
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
<b>Total</b>	18						

953

954