Toxicology Letters 416 (2026) 111813

Contents lists available at ScienceDirect

Toxicology Letters

journal homepage: www.journals.elsevier.com/toxicology-letters

Predicting acute developmental toxicity of chemicals in embryos of the African clawed frog (*Xenopus laevis*): Calibration and validation of regression-based quantitative structure activity relationship models for hazard assessment of chemicals in anuran amphibians

Christian Novello [a,1], Matteo Riccardo Di Nicola [b,1,*], Jean Lou CM Dorne [c], Erika Colombo [a], Edoardo Luca Viganò [a], Manuel E. Ortiz-Santaliestra [d], Nynke Kramer [e], Edoardo Carnesecchi [f], Anna Melina Steinbach [g], Ivano Eberini [h], Antony Williams [i], Emilio Benfenati [a], Alessandra Roncaglioni [a]

[a] *Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, Milano 20156, Italy*
[b] *Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d'Aosta, Via Bologna 148, Torino 10154, Italy*
[c] *Chief Scientist Office, European Food Safety Authority (EFSA), Via Carlo Magno 1A, Parma 43126, Italy*
[d] *Instituto de Investigación en Recursos Cinegéticos (IREC) UCLM-CSIC-JCCM, de Toledo 12, Ronda, Ciudad Real 13005, Spain*
[e] *Toxicology Division, Wageningen University, PO Box 8000, Wageningen 6700 EA, the Netherlands*
[f] *iDATA Unit, European Food Safety Authority (EFSA), Via Carlo Magno 1A, Parma 43126, Italy*
[g] *German Federal Institute for Risk Assessment (BfR), Max-Dohrn-Straße 8-10, Berlin 10589, Germany*
[h] *University of Milan, Dipartimento di Scienze farmacologiche e biomolecolari, Via Balzaretti 9, Milan 20133, Italy*
[i] *Center for Computational Toxicology and Exposure, Office of Research and Development, US Environmental Protection Agency (US EPA), Research Triangle Park, Raleigh, USA*

ARTICLE INFO

Editor: Angela Mally

Keywords:
Acute developmental toxicity, *Xenopus laevis* embryos
K-NN
Multiple regression models, QSAR

ABSTRACT

Global decline of amphibian populations has been correlated with a range of endogenous and exogenous variables including their unique physiology and ecology, exposure to chemicals, habitat reduction, climate change, as well as biological hazards such as emerging infectious diseases. The African clawed frog (*Xenopus laevis*) is an OECD test species used in toxicity testing as a specific proxy for humans and environmentally relevant species, for which acute toxicity data for a range of chemicals have been generated historically by industry, a number of public health agencies and academia. Of particular relevance are mechanistic effects of endocrine-active substances on metamorphosis and the thyroid axis, resulting in developmental toxicity. From such toxicity data, no open-source quantitative structure-activity relationships (QSARs) have been developed as in silico tools to predict such toxicity for data-poor chemicals in *X. laevis.* Such QSAR models can provide a quantitative starting point for the hazard assessment of chemicals in other anuran amphibians. This manuscript provides a description of the data collection and curation from the largest historical databases including the US EPA ECOTOX knowledgebase and the Ortiz-Santaliestra databases available for *Xenopus* embryos as acute median lethal concentrations (LC_{50}-12 h) for a total of 349 unique structures and 1978 individual entries. After data curation, the database contained 359 individual entries for a total of 175 compounds, and were computed using the negative logarithm of molar concentrations expressed as 12 h log 1/LC50 mmol/L. Subsequently, the database was then split into training set, test set and prediction set with 120, 40 and 13 compounds, respectively. These datasets were then used for the development and validation of two different QSAR models: 1. A k-Nearest Neighbours (k-NN) models using istKNN (in silico tools – KNN). 2. A multiple linear regression model (MLR) using the QSARINS (QSAR-INSUBRIA) software version 2.2.4. Overall, the QSAR models performed well for predicting acute toxicity of chemicals in *Xenopus* embryos and the MLR model performed slightly better than the k-NN model with correlation coefficients of 0.76 and 0.75 and root mean square errors of 0.63 and 0.67, respectively. However, underestimation of predictions for highly toxic compounds were observed and these

* Corresponding author.
E-mail address: matteoriccardo.dinicola@izsplv.it (M.R. Di Nicola).
[1] These authors share first authorship

https://doi.org/10.1016/j.toxlet.2025.111813
Received 28 October 2025; Received in revised form 20 December 2025; Accepted 29 December 2025
Available online 1 January 2026

limitations are discussed for both the k-NN and multiple linear regression model in the light of mechanistic interpretation and expert knowledge. Variability in the experimental datasets as well as under-representation of the most toxic compounds in the database are highlighted as major drivers influencing such underpredictions. Future directions from the present work include the modelling of other endpoints and developmental stages as well as other amphibian species using the available, although limited, data. Overall, it can be foreseen in the near future that such databases and models will be important to develop more performant in silico models, and ultimately to develop NAMs for ecotoxicity assessment of chemicals in anuran amphibians while reducing animal testing.

1. Introduction

Over the last three decades, efforts in computational toxicology research has been driving the development of *in silico* tools including (quantitative) structure-activity relationship ((Q)SAR), grouping approaches for setting chemical categories, assessment groups, expert systems, and read-across techniques. Such tools provide a means to predict toxicity properties of untested chemicals in a range of species of relevance to environmental risk assessment (ERA) (Astuto et al., 2022; Carnesecchi et al., 2020; Cattaneo et al., 2023; Lavado et al., 2021; Raunio, 2011; Toma et al., 2021).

QSAR models provide statistical relationships between physicochemical properties, structural characteristics, and toxicological effects of substances while supporting and depicting the mechanistic basis of toxicity and adverse outcomes in organisms (Cronin and Madden, 2010). Importantly, their development takes advantage of the availability of toxicological data stored in appropriate databases (e.g., the EFSA OpenFoodTox – Dorne et al., 2021).

Such approaches are central to the 21st century toxicology vision and next generation risk assessment approaches advocating the implementation of the 3Rs principle, i.e., the resolution to avoid animal experimentation altogether (Replacement), limit the number of animals (Reduction) and their suffering (Refinement) in tests to an absolute minimum (Russell and Burch, 1992). This may be attained through the integration of information coming from testing and non-testing methods within a weight-of-evidence framework as part of New Approach Methodologies (NAMs) (Berg et al., 2011; Bhattacharya et al., 2011; Casati, 2018; Di Nicola et al., 2023; National Research Council, 2007).

Key applications of QSAR models include prioritisation of industrial chemicals from a hazard perspective, classification and labelling, hazard identification during the first stages of R&D process as well as hazard characterisation through the generation of quantitative hazard metrics for human health, animal health and ecological risk assessment or screening of large ecotoxicological datasets (Astuto et al., 2022; Raies and Bajic, 2016).

Amongst all computational methods, QSARs currently represent one method accepted and promoted by regulatory agencies, academia, and industry as long as these meet the Organisation for Economic and Cooperation Development (OECD) requirements, including assessment of prediction accuracy through comparison with experimental data (Gozalbes and Vicente De Julián-Ortiz, 2018; OECD, 2014). In this context, the recently published OECD (Q)SAR Assessment Framework (QAF) provides harmonised principles and checklists for the assessment of QSAR models, individual predictions and results based on multiple predictions, with a strong emphasis on transparency and fitness for regulatory applications (Gissi et al., 2024; OECD, 2024). Amongst those QSAR models used for toxicity prediction, k-Nearest Neighbours (k-NN) models as non-parametric classifiers or regression estimators for continuous endpoints are probably the simplest ones (Altman, 1992). Overall, k-NN algorithms compute the outcome of a sample within a dataset based on the k most similar samples (neighbours) in a training set for which the toxicological outcomes are known from experimental datasets (Como et al., 2017; Manganaro et al., 2016). In other words, k-NN models allow for the prediction of continuous endpoints (e.g., lethal concentration (LC_{50})) as quantitative predictions of neighbours within a chemical space (Gadaleta et al., 2019; Gadaleta, 2014). A major advantage of k-NNs lies in the fact that they are easy to implement and often result in good predictive performance. In addition, the inspection of the selected neighbours allows one to apply a reasoning similar to the read-across approach. However, these require homogeneous features, are heavily dependent on k values and are sensitive to noisy data, missing values and outliers. Such k-NN models have been developed for a range of species relevant to ERA including rat, trout, honeybees and collembola to cite but a few (Como et al., 2017; Toropov et al., 2020).

Amongst organisms of interest to ERA, amphibians constitute an important taxonomical group and are recognised as the most threatened vertebrate taxa since they are declining in numbers and, over the last fifty years, hundreds of species have gone extinct. Such decline has been rationalised to be due to exposure to multiple stressors including habitat loss, climate change, chemicals and infectious diseases caused by fungi such as *Batrachochytrium* and viruses like *Ranavirus* (e.g., Campbell Grant et al., 2020; Falaschi et al., 2022; Fisher and Garner, 2020; Fisher et al., 2021; Green et al., 2020; Palomar et al., 2023; Rollins-Smith, 2020). In addition, their unique ecology, physiology, and life cycle with a range of aquatic (i.e., egg, embryo, tadpole) and terrestrial life stages (i.e., juvenile and adult) makes them particularly sensitive to such multiple stressors (Toropova et al., 2021; Toropov et al., 2022). The scientific panel of plant protection products and their residues (PPR) of the European Food Safety Authority (EFSA) has recently published a scientific opinion aiming to address the scientific basis of the sensitivity of amphibians and reptiles as well as data gaps and recommendations to integrate these taxa within pesticide ERA. The PPR panel highlighted that limited experimental toxicity data are available for amphibians and that there are currently very limited requirements to include them in pesticide ERA. Hence, the use of NAMs including *in silico* models such as QSARs and toxicokinetic-toxicodynamic (TK-TD) models provide practical tools to predict physicochemical properties, fate, and toxicity (see Di Nicola et al., 2023; Dorne et al., 2023; EFSA et al., 2018).

The development of QSAR models for tadpoles, as an aquatic phase of the amphibian life cycle, has been shown to be highly relevant since they may be particularly sensitive to chemical toxicity while undergoing metamorphosis (Gross et al., 2009; Zhang et al., 2019). Few QSAR models for tadpoles have been developed, so far, and these include models for a relatively limited number of alcohol compounds and species such as *Rana temporaria*, *Rana chensinensis*, and very recently for *Rana japonica* (Adhikari and Mishra, 2018; Agrawal et al., 2003; Huang et al., 2003; Jaiswal and Khadikar, 2004; Sahoo et al., 2016; Toropov et al., 2022; Toropov et al., 2023; Wang et al., 2018; Wang et al., 2019).

In this context, the recent regression-based *R. japonica* QSAR model (Toropov et al., 2022) highlighted the need to further develop QSAR models for other amphibian species for which some experimental data are available in public databases and the peer-reviewed literature (Toropov et al. 2022). Amongst amphibian species, the African clawed frog, *Xenopus laevis* (Daudin, 1802) (Fig. 1), an anuran amphibian from the *Pipidae* family, is of high relevance as it is a test species set by the OECD and much of the current knowledge of amphibian biology has been obtained using it as an experimental model. The reason why this species has been included as an OECD test species mostly lies in the fact that it is easy to keep in the laboratory being a sturdy and relatively small species with a life span as long as 30 years (compared to 5–15

Fig. 1. Adult individual of African clawed frog, *X. laevis* (Daudin, 1802). Photo credit: Matteo R. Di Nicola.

years in the wild), and its breeding habits generates a large number of eggs (i.e., a thousand eggs up to 3–4 times a year) (Cannatella and De Sá, 1993; Nikos, 2012; OECD, 2015; Reed, 2005). In addition, since molecular and cellular pathways are highly conserved, *X. laevis* has been used in developmental studies particularly for endocrine active substances, neurosciences, genetics and whole-organism-based drug discovery; indeed, in terms of evolution, it is closer to humans compared to other models, and its genome is well characterised, particularly in Xenbase (Blum and Ott, 2018; Dawid and Sargent, 1988; Straka and Simmers, 2012; Wheeler and Brändli, 2009). Finally, the ecotoxicological databases available in the peer-reviewed literature for *X. laevis* have been growing over the last few years, and these include the US EPA ECOTOX knowledgebase database and the recent Ortiz-Santaliestra et al. toxicological database submitted to EFSA (Ortiz-Santaliestra et al., 2017; 2018). It is particularly relevant to generate a predictive model accounting for general toxicity in amphibians particularly using Lethal Concentration (LC_{50}) since such an *in silico* model is not currently available for ERA in anuran amphibians (Ortiz-Santaliestra et al., 2018). This differs from *in silico* models based on effective concentration (EC_{50}) from the frog embryo teratogenesis assay in *Xenopus* (FETAX), which would address teratogenicity outcomes.

Acute toxicity in amphibian embryos arises from a range of molecular initiating events that impair essential physiological functions early in development. The most common MoA is baseline narcosis, a non-specific membrane perturbation caused by hydrophobic organic chemicals that accumulate in lipid bilayers and disrupt cellular homeostasis (Escher and Hermens, 2002; Verhaar et al., 1992). In addition, ionoregulation disruption is a well-recognised MoA in early amphibian development: ion-transporting cells in the embryonic and larval epidermis, together with Na+ /K+ -ATPase-dependent transport, contribute to osmotic balance; these processes can be impaired by metals and by surfactant-containing agrochemical formulations, leading to rapid mortality (Edginton et al., 2004; Freda, 1991; Quigley et al., 2011). Mitochondrial dysfunction and oxidative stress can contribute to acute embryotoxicity, as several stressors cause mitochondrial injury and increased reactive oxygen species, with downstream impairment of cellular energy homeostasis and survival (Carotenuto et al., 2022; Lushchak, 2011). Neurotoxicity may also play a role even at early developmental stages, particularly for acetylcholinesterase-inhibiting insecticides, which can disrupt cholinergic signalling and locomotor function in amphibian larvae (Sparling and Fellers, 2007). Endocrine-mediated mechanisms can contribute to developmental toxicity as well; in amphibians, thyroid hormone signalling is a key regulatory system and there is evidence that thyroid hormone related processes are required before, and not only during, metamorphosis, making them a plausible target for thyroid-active contaminants (Carr and Patiño, 2011; Tindall et al., 2007). Together, these MoAs illustrate that acute lethality in *X. laevis* embryos may reflect both baseline toxicity and more specific, non-baseline mechanisms, which can influence the structure-activity relationships captured by QSAR models.

Hence, the aim of the present study is to collect available LC_{50} values in *X. laevis* embryos from public databases and peer reviewed literature for a wide range of chemicals to develop QSAR models including k-NN and Multiple Linear Regression (MLR) models. In addition, one of the additional practical goals of the current study is to apply these QSAR models as valuable tools for assessing the association between structural and sub-structural features of chemicals and toxicological features in *Xenopus* to move towards the integration of NAM-based hazard identification and characterisation for this species, in the light of the holistic "One health" principle, particularly for data poor chemicals (FAO-OIE-WHO, 2019). Finally, this manuscript concludes on future research perspectives.

2. Methods

2.1. Data collection and curation of acute toxicity data for chemicals in embryos of Xenopus laevis

Data collection was performed from the US EPA ECOTOX knowledgebase database (https://cfpub.epa.gov/ecotox/) and the Ortiz-Santaliestra et al. ecotoxicological database submitted to EFSA (Ortiz-Santaliestra et al., 2017; Ortiz-Santaliestra et al., 2018). Data for 349 unique structures were initially retrieved and associated with a total of 1978 entries (see Table S1 for the types of information included in the database).

Most of the LC_{50} data were split according to major developmental stages (embryo, tadpole, hatchling, metamorphic, and juvenile). The most suitable stage to investigate was the embryo, approximately corresponding to stages from 1 to 20 of the Gosner classification (Gosner, 1960), since it is the most represented and because of its toxicological sensitivity to waterborne pollutants amongst amphibians across the different developmental stages (Gross et al., 2009; Zhang et al., 2019). Endpoint values associated to "> " or "< " qualifiers were excluded, as well as low-purity test substances such as plant protection products commercial formulations. The preliminary, non-stringent purity cut-off for including test substances in the database was set at 50 %. Among the substances above this threshold, very few had only moderate purity. For example, diazinon had 60 % purity and was placed in the prediction set, and only four other substances had purity below 90 %. The vast majority of substances showed a purity greater than 95 %.

Any entry presenting deviating data (duration of exposure, pH, and temperature parameters) with respect to the Standard Guide for Conducting the FETAX (ASTM, 1991) was excluded too.

The exposure route was dermal for all datasets. Exposure of embryos to waterborne contaminants is indeed typically simplified as dermal, as it occurs through a passive, surface-based uptake rather than an active ingestion. The experimental data was generated using de-jellied embryos, with a cysteine solution adjusted with NaOH, before putting them in contact with waterborne contaminants. Measurement units were standardised and endpoint values were converted from mg/L to mmol/L, by dividing by the molecular weight of the tested substance. Therefore, $Log_{10}LC_{50}$ values were obtained and used.

For compounds associated with multiple values, a thorough assessment was performed to identify those with high experimental variability. A factor of 4 was used, as proposed previously (Benfenati et al., 2007). Overall, the basis for such a factor relies on the fact that, if the maximum value (in mg/L) among the replicates is up to 4-fold higher than the minimum value, the chemical can be included and the average LC50 across replicates can be calculated (see Figures S1 and S2). Conversely, if the ratio between the maximum and the minimum value is higher than 4-fold, the chemical is set aside to the "prediction set" (Figure S3). This approach was applied here to exclude the noisiest data and to ensure that consistent datasets were used for QSAR model development. Most importantly, the main objective of the prediction set was to probe the model's performance under realistic and not overly-optimistic conditions, as well as to evaluate whether it could

produce useful predictions when experimental reproducibility is low.

Each individual compound reporting unique experimental values and homogeneous replicates were included in the database and was associated with a dedicated ID and its mean LC50 value. Entries were sorted in ascending order according to the $Log_{10}LC50$ or $Log_{10}LC50$ mean for compounds with reported multiple data points.

The final database (after the prediction set was excluded) was split into training and test sets in a 3:1 proportion (Figure S4). The split of the molecules in the two datasets was not guided by an *a priori* chemical similarity analysis but rather through a randomised approach. Indeed, the test set was randomly extracted while simultaneously ensuring that the compounds were similarly distributed in terms of toxicity values between the two sets. The two populations of substances in the training and test sets were assessed for their similarity. For this purpose, the distribution of the different classes of compounds in the two sets was checked by means of a Principal Components Analysis (PCA). The PCA on the molecular descriptors of the two datasets was performed with the purpose of clustering chemicals according to their distribution in the first two Principal Components (PC1 and PC2), and ultimately assessing that the chemical properties of the training set were well-represented in the randomly generated test set.

The molecular descriptors were calculated using the Dragon 7.0 application (Kode Chemoinformatics). A total of 707 2D molecular descriptors were selected from the list while all 3D molecular descriptors were disregarded since no geometrical optimisation was performed.

After all calculations, the correlated molecular descriptors were pruned; so that all descriptors with a correlation coefficient equal or greater than the selected threshold value of 0.95 were excluded.

2.2. Development of QSAR models

2.2.1. istKNN

k-Nearest Neighbours (k-NN) models were built using the istKNN (*in silico* tools – KNN) 0.9.3 software developed by Kode srl based on Chemistry Developmental Kit (CDK) version 1.4.9 and VEGA core libraries (https://www.vegahub.eu/) version 1.1.2. istKNN implements a k-NN algorithm, which adopts a similarity-based approach (Altman, 1992; Como et al., 2017; Friel and Pettitt, 2011; Gadaleta et al., 2019; Gadaleta, 2014; Manganaro et al., 2016).

The similarity amongst chemicals has been described through an integrated similarity index (SI), developed inside the VEGA platform. The SI ranges from 1 (maximum similarity) to 0 (minimum similarity) and results from the weighted combination of fingerprints with non-binary structural keys based on constitutional molecular descriptors (Floris et al., 2014).

The istKNN in silico tool allows for refinement of the classical k-NN algorithm by setting additional conditions that a target chemical should fulfil to be considered reliably predicted. Indeed, the k nearest neighbours used for prediction should have a similarity value with the target greater than a given threshold (T_{sim1}), otherwise they would not be used for prediction. If no neighbour matches the threshold, the model does not provide any prediction for the target compound (missing value). If only one neighbour matches the threshold, the similarity should be higher than a second stricter threshold (specifically named "similarity threshold for single molecule") to yield a prediction (which corresponds to the experimental value of this selected neighbour, in this case). If two or more neighbours fulfil the T_{sim1}, the range of experimental values of the identified neighbours is considered. If the difference between the maximum and minimum experimental values of the neighbours is lower than a threshold ($T_{min–max}$), the target is predicted, otherwise the model does not provide predictions. To calculate the prediction when more than one neighbour is selected, the experimental values of the similar compounds are weighted differently based on their similarity with the target (by setting an enhancement factor that increases the weight of the most similar compounds in the prediction) (Gadaleta et al., 2019).

The robustness of each model is then assessed using Leave-One-Out (LOO) cross-validation (Gadaleta, 2014).

Table 1
Ranges of values and steps chosen for the five istKNN customisable parameters.

Parameter	From	To	Step
Number of Neighbours (K)	2	5	
Similarity Threshold	0.65	0.95	0.05
Similarity Threshold for single molecule	0.8	1	0.05
Enhancement factor for Weights	1	5	
Experimental Range	0.5	4	0.5

The Batch mode allows the user to explore several possible parameter settings to automatically produce many models on the training set. Each model is defined by a special combination of values relative to the five customisable parameters, and the user can go through the output to select the optimal combination (Como et al., 2017).

k-NN was used as a regression estimator to develop batch models according to the parameters' ranges and steps shown in Table 1.

The most suitable combination of variables for the model was selected with a view to balance R2 with the number of non-predictions, and ultimately to make a compromise between accuracy of the model and "inclusivity".

To do so, a univariate analysis was performed on three parameters: the number of neighbours, the percentage of unpredicted compounds and the minimum similarity. Each value or range of values that they could take was singularly plotted against the best R^2 value they were associated to (Figures S5-S7). The one related to the highest R^2 or to a rapid R^2 increase was selected. The model was fitted on the training set using the software's Build mode so that the model run on the test set to assess its capacity to make predictions on an external dataset, and on the prediction set, for benchmark purposes. Statistical analyses were carried out for each of the model's performances to identify the missing values and the outlier chemicals.

2.2.2. QSARINS

QSARINS (QSAR-INSUBRIA) software version 2.2.4 was used for the development and validation of Multiple Linear Regression (MLR) models. It includes tools for modelling datasets and exploring their chemical space by PCA, based on externally calculated descriptors (Gramatica et al., 2013; Gramatica et al., 2014; QSARINS, 2019).

A total of 1409 PaDEL molecular descriptors were calculated with PaDEL-Descriptor (version 2.21) (Yap, 2011) and imported into QSARINS as independent variables for the chemicals, which were already subdivided into training and test sets, while the experimental values were set as dependent variables. A pre-reduction of input molecular descriptors was performed to mitigate the redundancy of the inter-correlated ones giving similar structural information. Descriptors to disregard were automatically identified by calculating pairwise correlations so that the correlation amongst all pairs of descriptors was computed. As a consequence, when a pair was found to be highly correlated (correlation higher than the default cut-off value of 95 %, as in Dragon), the descriptor with the highest correlation compared to other descriptors was removed. The chemical space of the dataset was inspected by means of PCA and both the score and the loading plots were scrutinised to confirm the molecules' clustering seen from the PCA performed in Dragon 7.0.

MLR models were then developed in QSARINS using the Ordinary Least Squares method (OLS). For variable selection, a Genetic Algorithm (GA) procedure was applied (Gramatica et al., 2013) to identify the optimal combination of descriptors in high dimension models, instead of exploring all combinations of small dimension ones.

Fine tuning of the GA was performed while setting the size of model populations equal to 700, and a total of 2000 generations (i.e., reiterated processes to isolate the best performing models). The algorithm was run twice, first introducing a 20 % mutation rate (i.e., rate of descriptor substitution) and then an 80 % one. Typically, if both times the

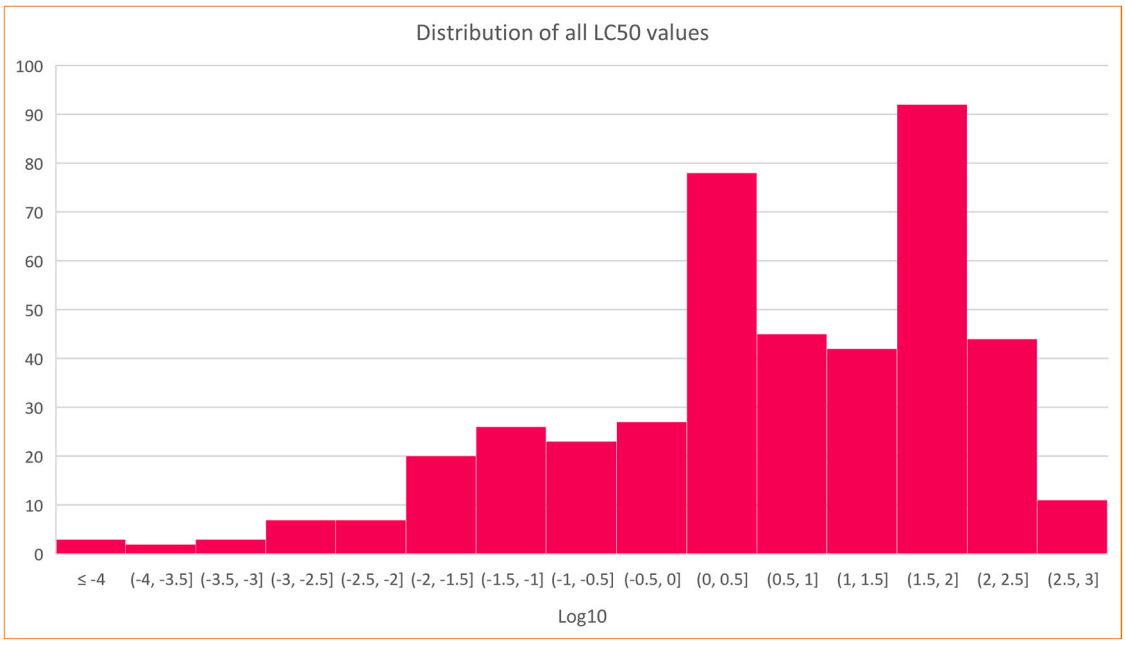

Fig. 2. Distribution of acute mortality values in *X. laevis* embryos expressed as $Log_{10}LC50$ (mmol/L).

algorithm converges to the same output, the generated models can be reasonably considered an optimal solution.

A maximum number of possible variables was also set in order to avoid overfitting of the model (usually the number of variables must be 1/5 or 1/7 of the total number of objects).

During GA selection, a big batch of models was obtained in around two days of computation.

QSARINS was used to check the calculated models' validity by fitting and by both internally and externally validating them, as described in 2.3. The best variables' combination was selected from the list considering all criteria, but especially the external ones, as detailed in the Results section.

2.3. Statistical analysis

The performance of k-NN predictions was evaluated using fitness metrics expressed as a regression coefficient (R^2) and RMSE, which are calculated on both the training predictions using LOO cross-validation and predictions of external datasets. k-NN models were further assessed for their predictive capacity using the ratio of compounds

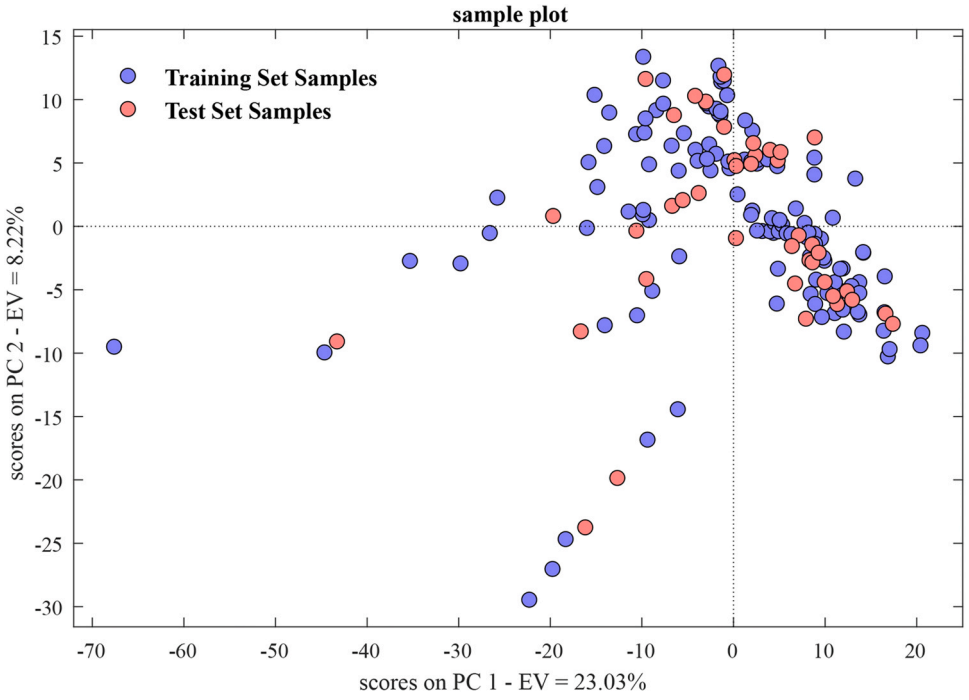

Fig. 3. Score scatter plot of the principal component analysis (PCA) model performed on acute LC_{50} values (LC_{50} 12 h) in *X. laevis* embryos with the MATLAB PCA Toolbox (Ballabio, 2015). The two axes represent two Principal Components (PCs) and the corresponding percentage of explained variance (%EV). Compounds are distributed according to the correlation of the original variables with the PCs which can attribute a given %EV of the chemical data variability.

Table 2
Settings of the k-NN model for predicting acute toxicity of chemicals in *X. laevis* embryos (LC_{50} 12 h).

k (neighbors' number)	4
Similarity threshold	0.65
Similarity threshold for single molecules	0.80
Enhancement factor	5
Experimental range	2.5

correctly predicted compared to the total number of compounds in the database.

The fitting of MLR models was also evaluated using the R^2, the RMSE and a modified R^2 form, R^2_{adj}. R^2_{adj} also assesses the models' degrees of freedom, namely the convenience of adding a new descriptor to it. The robustness of each model was evaluated by the Cross-Validated coefficients of determination R^2 and Q^2, using LOO cross-validation.

Internal validation was also performed by the stronger LMO technique (R^2_{LMO} and Q^2_{LMO} were calculated). Lastly, to demonstrate that the models were not the result of chance correlation, the Y-scrambling procedure was applied. Visual inspection for selecting the best MLR model took advantage of the creation of different plots within QSARINS, including the plot of the experimental vs. predicted activity (Fig. 6), the Williams plot (Figures 7 and S10), the Residuals (Fig. 8), the Leave-Many-Out (LMO) (Figure S8) and the Y-scramble plots (Figure S9).

3. Results

3.1. *Xenopus* embryo database

The *Xenopus* embryo database resulted in 430 individual LC_{50} values for a total of 175 compounds. The distribution of Log LC_{50} values from the *Xenopus* embryo database is illustrated in Fig. 2 and highlights the variability in the toxicity values of the database.

Amongst the 175 compounds, 122 provided a unique experimental value and used for the training set while 40 and 13 compounds were used for the test and prediction sets, respectively. To note that both the test and prediction sets contain multiple individual values (up to 23). The training and test sets were generated by splitting the LC_{50} values for the 122 and 40, respectively, and were associated with 359 unique experimental values or multiple consistent values (Figures S1-S4).

The results of the PCA analysis performed on Dragon descriptors are displayed in a Score plot (Fig. 3), showing that the samples of the training and test sets projected in a similar manner for the relevant PCs. Just a few outliers were noted as noted on Fig. 3 in the far left and bottom of the plot. These outliers were interpreted as not being of statistical relevance since the training set is 3-fold larger than the test set.

As the two data clouds mostly overlapped in the orthogonal space, it was concluded that the randomly-generated test set was structurally similar to the training set, and therefore as a representative subset of data.

As discussed in the materials and methods section, the less reliable

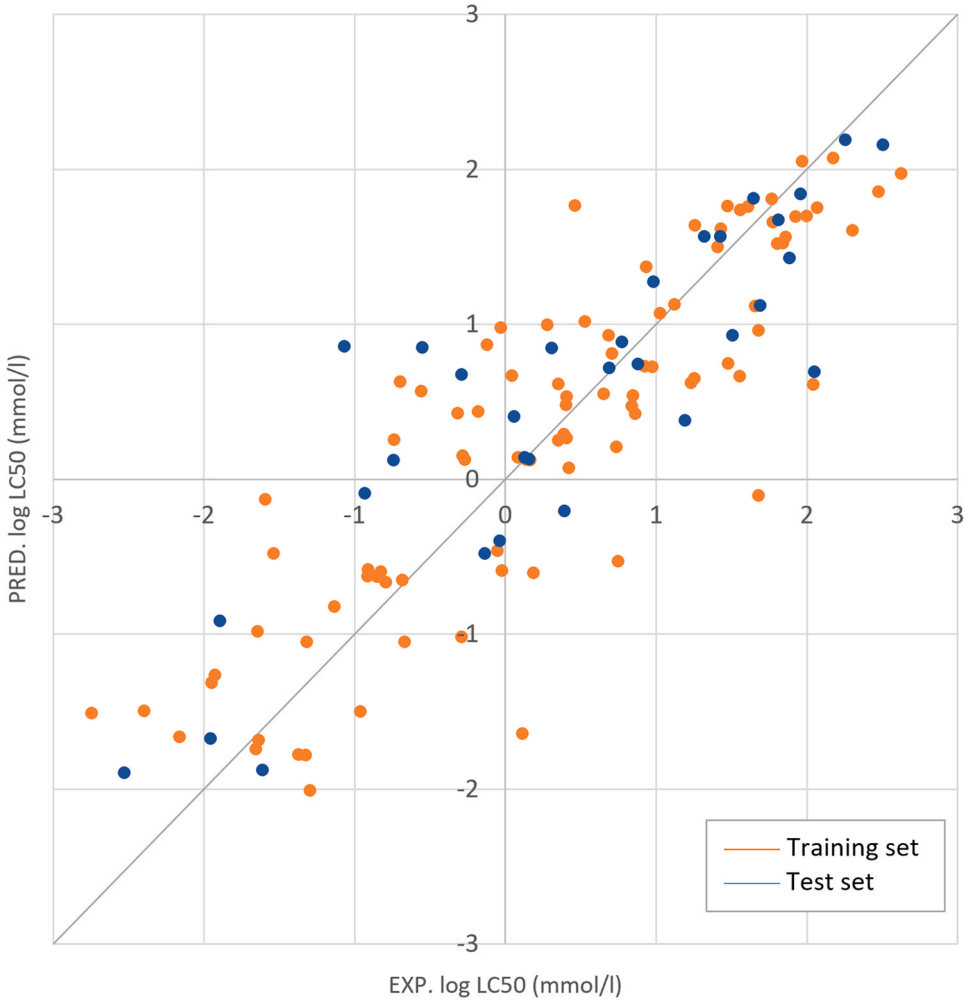

Fig. 4. Overlapping experimental vs. predicted training and test sets for acute toxicity in *X. laevis* embryos (LC_{50} 12 h). Acute toxicity are expressed as $Log_{10}LC50$ (mmol/L). Training and test sets are represented in orange and blue, respectively, and trendline is drawn in grey.

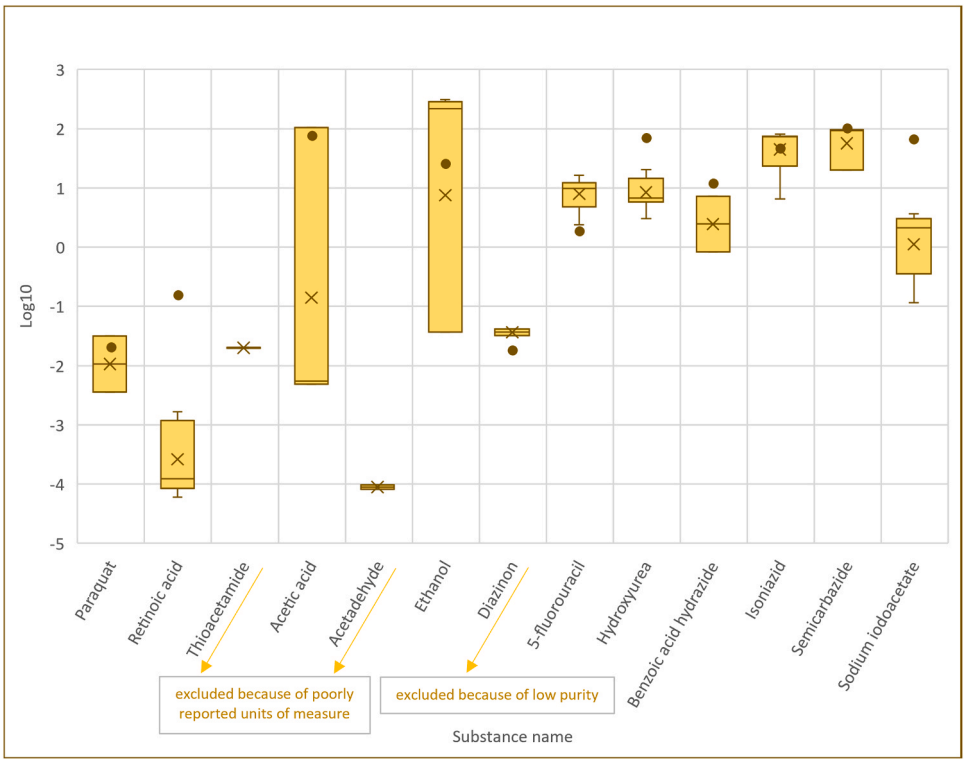

Fig. 5. Comparison of the predicted values (represented as brown points) for acute toxicity in *X. laevis* embryos (LC_{50} 12 h) for the prediction set by the k-NN QSAR model as compared with box and whisker plots of experimental values (in yellow). Acute toxicity values are expressed as $Log_{10}LC50$ (mmol/L). Exclusion criterion for the substances is reflected as high variability in experimental values, if not indicated otherwise.

experimental data were included in the prediction set, while the training set contained more reproducible data. Among these, three chemicals were added to the prediction set for a different reason than data variability including diazinon with low purity (60 %), and acetaldehyde and thioacetamide because of alleged incongruities in the units of measure from the reference studies. Indeed, the data points associated with these compounds are associated with a study (Fort et al., 2003) which in some cases showed important inconsistencies with other homologous studies (see Dawson and Bantle, 1987; Dawson et al., 1996; Dresser et al., 1992; Fort et al., 2004). Data curation resulted in excluding acetaldehyde and thioacetamide from the prediction set since units of measure were poorly reported and may have resulted from a switch between % w/v and mg/L and this is underlined by a 4 order of magnitude difference. Overall, the prediction set consisted of a set of 13 compounds.

3.2. k-NN model performance

The k-NN model output resulted in 5600 combinations for all values using the five parameters inherent to the model namely k, Min Similarity, Min Similarity for a single molecule, Enhancement factor and Experimental range as illustrated in Table 2. Each combination was associated with the number of valid predictions so that the model was able to provide both the number and percentage of unpredicted values.

As expected, the minimum similarity and the number of valid predictions were inversely related; as the former increased, the latter decreased dramatically (from a maximum of 114 to a minimum of 16, corresponding to a percentage of change in unpredicted values change ranging from 7 % to 87 %. The higher the similarity value required to include neighbours, the higher the number of missing values so that no compounds were similar enough to produce a prediction. Parameters characterising the best model ensured an optimal balance between stringent conditions and coverage are summarised in Table 2.

3.2.1. Model performance

The k-NN model was built based on the training set and its predictive capability was tested as follows. From the 122 chemicals in the training set the model was able to perform 93 valid predictions in LOO mode (76.2 % of the dataset). The R^2 was 0.746 and the RMSE was 0.63. 10 outliers, out of 93 predictions, were noted. The outliers were defined as the chemicals whose predictions differed by more than 1 unit (in terms of Log10LC50 in mmol/L) from the experimental values. The remaining 29 compounds (23.8 % of the dataset) were not predicted, due to two possible reasons: 1. For 21 out of 29 compounds, the experimental range of similar molecules exceeded the given threshold and all were associated with 4 neighbours with an experimental variability of at least 2.5 orders of magnitude. 2. For the remaining 8 missing values, none, or at best one similar molecule was found to able to provide a prediction (Fig. 4). Such statistics were also calculated while running the model on the test set. The algorithm was able to carry out a total of 32 valid predictions on the test set, representing 80 % of all values, with a coefficient of determination of 0.74 and a RMSE of 0.67. Among the predicted compounds, 3 response outliers were noted so that 20 % of the dataset representing 8 compounds remained unpredicted. Similarly, most of the missing values had 4 neighbours whose experimental variability exceeded the cut-off value of 2.5. Only 1 compound was not predicted because of the total lack of suitable molecules (Fig. 4).

For both experimental and predicted datasets, most of missing predictions were associated with group of compounds with the highest observed toxicity. Indeed, half of the unpredicted compounds, namely 15 values in the training set and 4 in the test set, had an experimental $Log_{10}LC50$ lower than −1, even though compounds with such potency accounted for 25 % of the training set (30 out of 122) and 20 % of the test set (8out of 40). The model was also assessed while investigating its performance on the prediction set particularly the impact of the different exclusion criteria. Overall, the k-NN model performance was reliable since toxicity predictions provided sound results for 11 compounds out of a total of 13. Indeed, acetaldehyde and thioacetamide

Table 3
Fitting, internal and external criteria and associated parameters for the best QSARINS model predicting acute toxicity of chemicals in *X. laevis* embryos (Lethal Concentrations 12 h (LC_{50}-12 h).

Fitting criteria	R^2	0.7637
	R^2_{adj}	0.7514
Internal validation criteria	Q^2_{loo}	0.7379
	Q^2_{LMO}	0.7269
External validation criteria	$RMSE_{ext}$	0.6302
	Q^2-F3	0.7919

were disregarded (15.4 % of the dataset), due to the high experimental variability of the 4 identified neighbours.

In this case, predicted values were not set against the mean LC_{50} for each compound, but rather their box and whisker representations providing a graphical representation of predictions with respect to variability and uncertainty in the distribution of values. Most of the compounds in the prediction set were excluded from the main analysis because of their variability (Fig. 5).

This comparative analysis highlighted that almost all predictions were close to the range of experimental values with the exception of retinoic acid and sodium iodoacetate.

3.3. QSARINS model

3.3.1. Best model choice

Optimisation of the model with regard to the number of variables was performed to optimise the outcome between predictivity (low number of variables) and descriptiveness (high number of variables). The models with an optimised number of 6–7 variables were classified using their Q^2_{F3} (Consonni et al., 2010). Such classification also took into account parameters of internal validation, such as Q^2_{loo}. In addition, models with low squared-correlation coefficient values ($Q^2_{loo} < 0.7$, according to Gramatica, 2007) indicated low robustness and low internal predictive ability and the fitting of parameters, such as R^2_{adj}, were also considered. Table 3 provides the metrics of the best QSARIN model.

The predictivity of the QSARIN model was also assessed through an analysis of the changes in metrics following the chemicals' distribution across the training and test sets. Given that no rearrangement led to significant variations in the model's parameters, it was concluded that prediction performance was good and not associated with the random optimal allocation of the samples to the training or test sets (data not shown).

3.3.2. Model features

The experimental vs. predicted (obtained by model equation) scatter plot for both the training and test sets is displayed in Fig. 6.

Overall, each of the 162 compounds is associated with a predicted value while distribution of the data points indicates outlying predicted

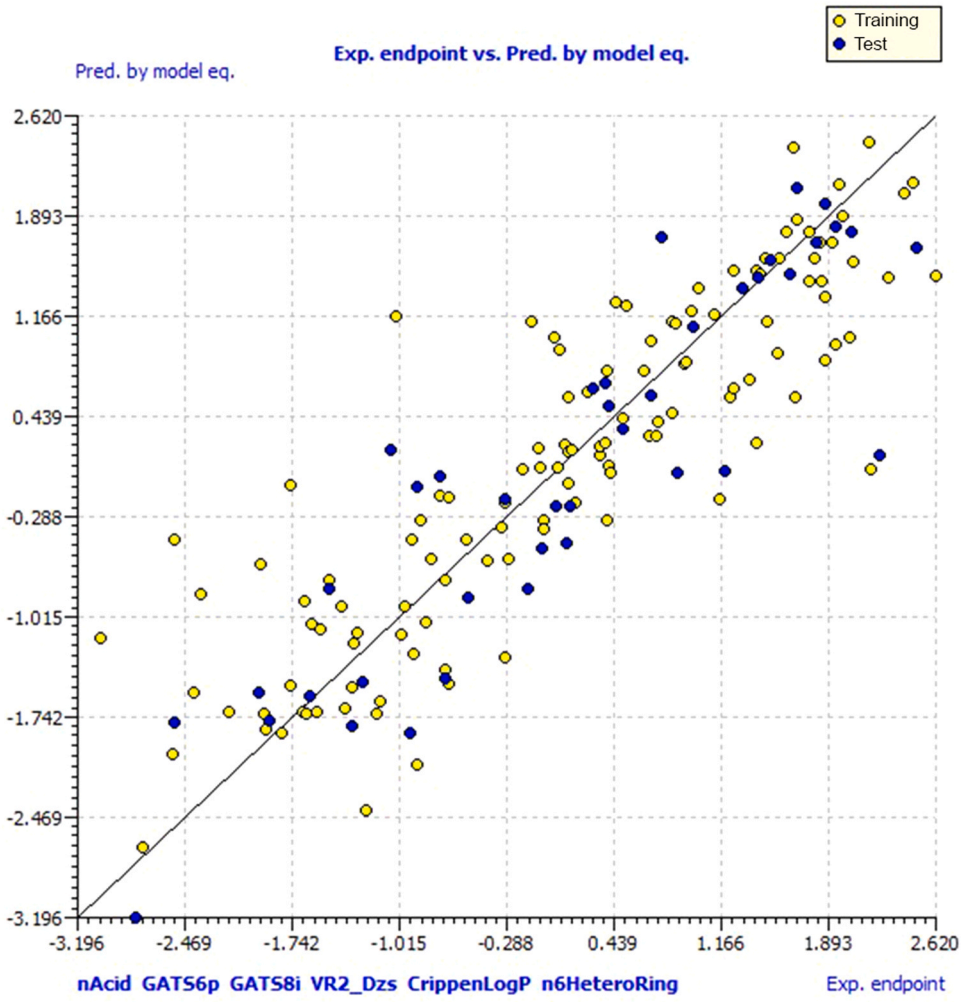

Fig. 6. Overlapping experimental vs. predicted training and test sets for acute toxicity in *X. laevis* embryos (LC_{50} 12 h). Acute toxicity is expressed as $Log_{10}LC50$ (mmol/L). Training and test sets are represented in yellow and blue, respectively, and trendline is drawn in black. Model variables are listed below the x-axis.

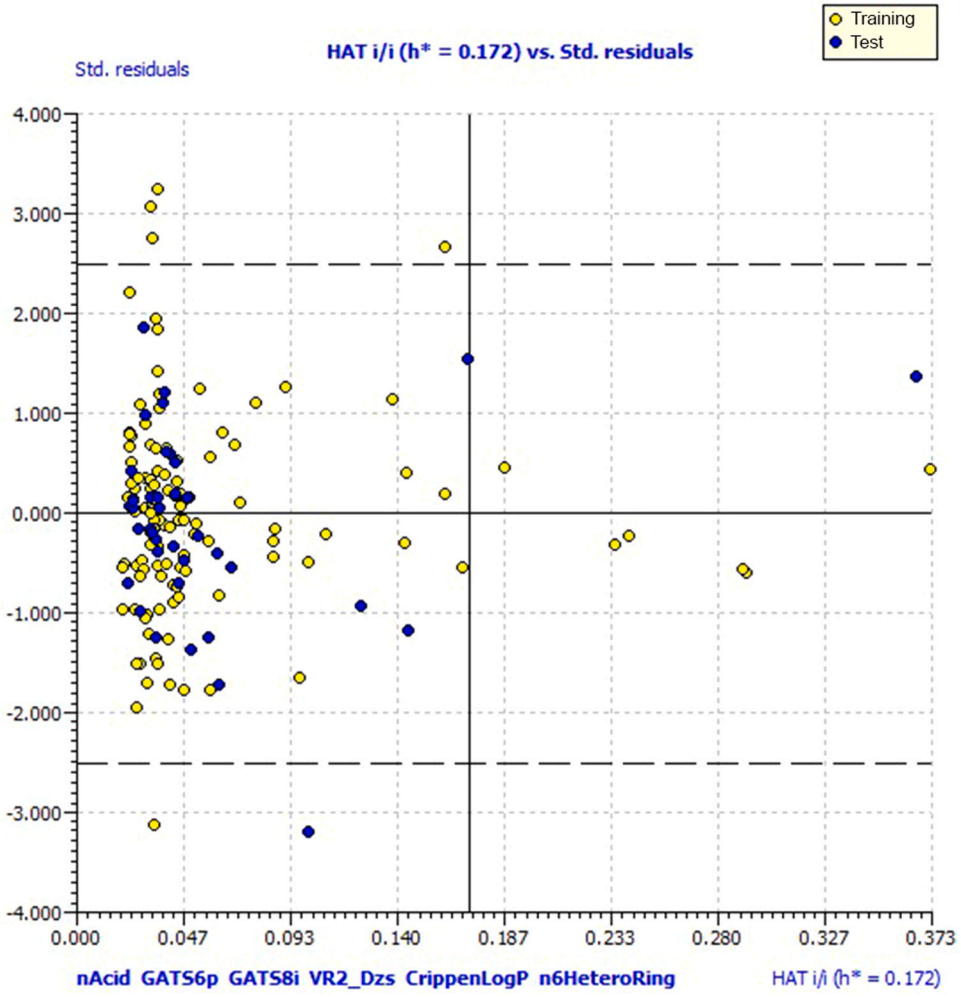

Fig. 7. Williams plots describing overlapping experimental vs. predicted training and test sets for acute toxicity in *X. laevis* embryos (LC_{50} 12 h). Training and test sets are represented in yellow and blue, respectively. The vertical line corresponds to the HAT threshold value of the structural domain (h*), while the dashed horizontal lines are the user-defined thresholds for Y-outliers. The model's variables are listed below the x-axis.

values with respect to the experimental values. Identification of such outliers was further validated via residuals analysis, as discussed below.

According to the tiered approach proposed by Hanser et al. (2016), the concept of decision domain of a predictive model must be based on three independent steps, namely the Applicability Domain (AD), the Reliability Domain (RD) and the Decidability Domain (DD). In the current assessment, particular attention was paid to the AD and RD. First, the AD was determined with a leverage approach by means of a Williams plot (Fig. 7). In this case, 8 molecules fell outside the AD (data points on the right of the critical leverage h* vertical line, in Fig. 7).

To define the RD, the same principle applied to the k-NN model was adopted for the QSARIN model for consistency. The prediction outliers from the k-NN model were associated with compounds for which predicted values differed by more than 1 unit from the experimental values (in terms of $Log_{10}LC50$ in mmol/L).

The Residuals plot supports visualisation of the outliers, with residual values higher than one, and therefore the most distant from the central line (Fig. 8). Overall, a total of 20 predictions (out of 162) were identified as falling outside of the RD. In addition, the goodness of the residuals' distribution was also assessed. Ideally, residual plots should be symmetrical with respect to the origin and should have a high density of points close to the origin and a low density of points away from it. Most importantly, their residuals should be independent and normally distributed, i.e., there should not be any clear pattern or trend in the residuals' distribution, since this would mean that an unknown determining factor was not considered and that the model still requires improvement. In this case, the Residual plot met all these requirements (Fig. 8).

4. Discussion

4.1. k-NN model

The performance of the k-NN model for the prediction of acute toxicity of chemicals in Xenopus embryos has been described above and provided satisfactory results with R^2 and RMSE of 0.75 and 0.63, respectively. Comparison between experimental and prediction results with regard to acute toxicity of chemicals in *Xenopus* embryos for both the training and test sets revealed the presence of patterns in the distribution of the data points. First, the data cloud "flattened" in correlation to the intermediate values. For example, it is clear that points aligned themselves horizontally in the space between −1 and 1 of the x-axis. In fact, in this case the model formulated the same prediction (same y-value) for chemicals with observed values differing by up to 3 orders of magnitude (Fig. 9).

Due to the tendency of the model to provide mean predictions predict in the middle", it was observed that the lethality of the most potent compounds was often underestimated (data points on the left of the plot were above the trendline, Fig. 10); conversely, the ones associated with the least toxic compounds were usually overestimated (data points on

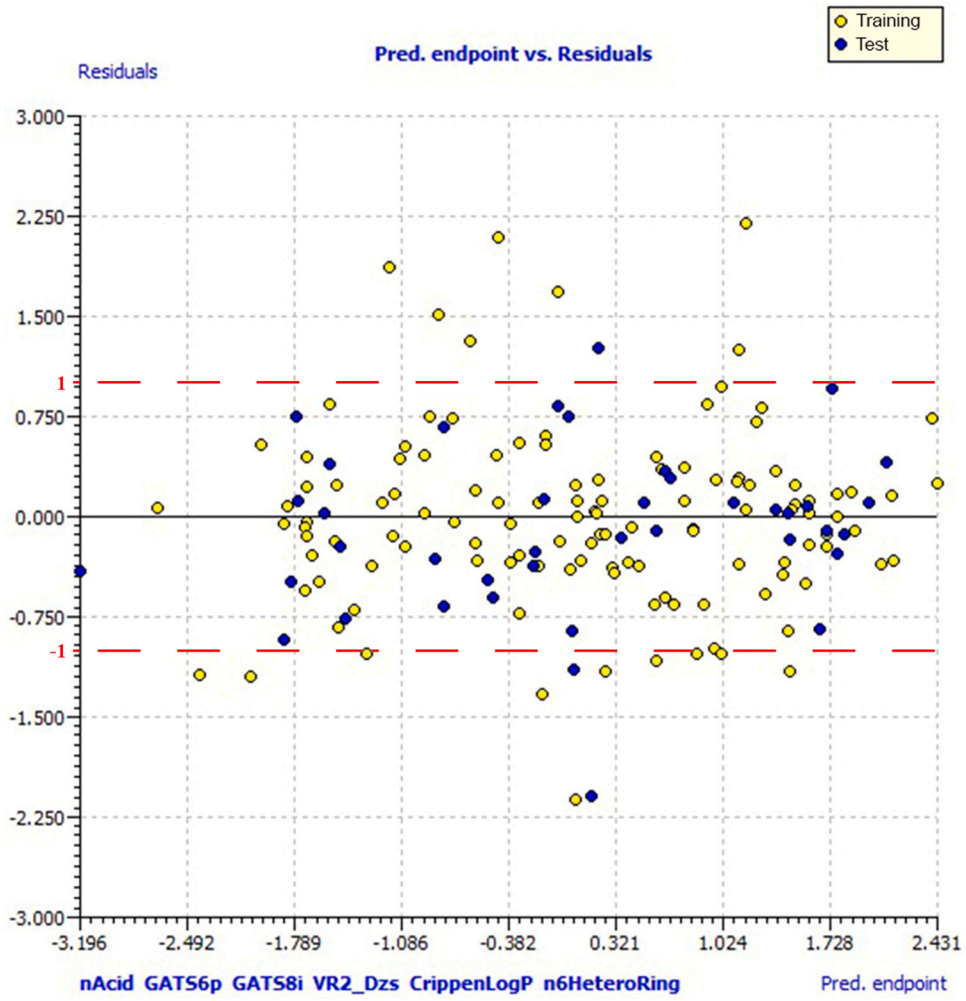

Fig. 8. Overlapping Residual plots describing overlapping experimental vs. predicted training and test sets for acute toxicity in *X. laevis* embryos (LC_{50} 12 h). Training and test sets are represented in yellow and blue, respectively. Predicted acute toxicity values are expressed as $Log_{10}LC50$ (mmol/L). The model's variables are listed below the x-axis.

the right of the plot were below the trendline, Fig. 10).

Overall, underestimating toxicity can have consequences in hazard identification and characterisation, compared to overestimation which would be more conservative; thus, for such potent compounds, model performance is an issue to which particular attention should be paid. Variability and uncertainty in experimental data as well as reliability of such datasets is also an issue that should also be mentioned and a weight of evidence approach should be applied to provide a transparent assessment of both in silico and experimental results.

Overall, given the way the data cloud narrowed with increasing LC50 values, it can be concluded that the k-NN model is less accurate at predicting highly acute toxic compounds. This situation may be related to the variability in the data distribution described at the beginning of the results section, particularly because the most potent compounds severely underrepresent the database when compared with most other compounds. Thus, the most potent substances were most likely not statistically significant and relied on similarly toxic compounds as called named "attractors" of the model and in the difficulty to form a consistent body of predictions. In most cases, the algorithm did not provide a prediction because of the high variability of the neighbours' observed toxicity values. Thus, a consistent number of failed predictions are attributable to the stringent experimental range cut-off. Finally, it should also be taken into consideration that predictions are based on 4 neighbours and that, as the number of similar compounds increases, so does the probability of the predicted endpoint values to diverge.

4.2. QSARINS modelling

The evaluation of the QSARINS prediction results mostly confirmed the prediction results from the k-NN models, even though the QSARINS model performed slightly better (R^2 and $RMSE_{ext}$ of the QSARINS model was equal to 0.76 and 0.63, respectively, as reported in Table 3, while R^2 and RMSE of the k-NN model on the test set was equal to 0.75 and 0.67, respectively).

The data point distribution for the QSARINS model's as illustrated in the plot comparing experimental vs. predicted values (Fig. 6) also reflected this conclusion from the results of the k-NN model (Fig. 4). A similar tendency of the data cloud to be more scattered in the region of the most potent endpoint values was observed, as well as the horizontal alignments in the middle of the distribution, and the overestimation of the toxicity of the compounds associated with the highest LC50 values. Most importantly, a correlation was found between the unpredicted and outlier compounds of the k-NN model and the chemicals excluded from the Applicability and Reliability Domain of the QSARINS model.

From the interpretation of the Williams plot, it is possible to rule out 8 compounds from the Applicability Domain: Methotrexate, Remazol Turquoise Blue, Maneb, Acetone, Dimethyl sulfoxide, Alpha-chaconine, Alpha-solanine, and Pentachlorophenol, whose leverage values were so high that they could not be detected in the original graph (Figure S10). Most of the chemicals that were out of the Applicability Domain (6/8) were also not predicted by the k-NN model. Among them, 4 were not

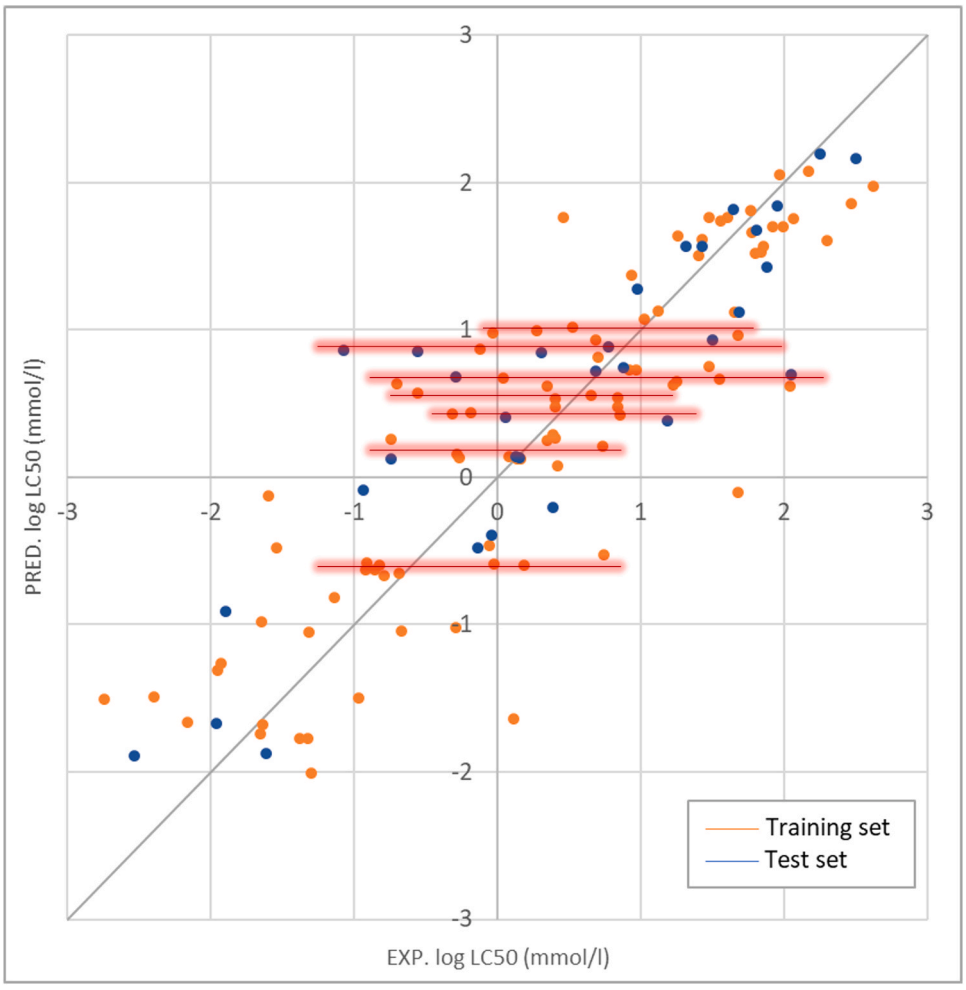

Fig. 9. Superimposed plots describing overlapping experimental vs. predicted training and test sets for acute toxicity in *X. laevis* embryos (LC_{50} 12 h). Training and test sets are represented in orange and blue, respectively. The red lines highlight where data points align, due to the tendency of the algorithm to assign the same intermediate value to different predictions.

predicted because of the lack of similar compounds in the database.

With regard to Y-outliers, compounds with highly-deviating standardised residuals were derived from the Williams plot. To define the Reliability Domain more stringent conditions were applied, in order to be consistent with the k-NN model's definition of outliers, as previously described, and resulted in raising the number of outliers to 20. Most of these 20 compounds were also not associated with valid k-NN predictions with 11 molecules failing to predict namely 8 compounds due to high experimental variability of similar chemicals and 3 due to the absence of suitable neighbours as well as 3 outliers.

4.3. Chemical space coverage and models applicability

The present analysis demonstrates that the developed QSAR models for acute toxicity in *Xenopus laevis* embryos perform reliably within the chemical space represented by the curated dataset, providing robust predictions for a fairly broad range of compounds. An examination of the chemical space highlights that highly toxic, non-baseline toxicants, including highly electrophilic and reactive substances, strong acids and bases, and organometallic species, are currently underrepresented. These compounds act through specific molecular initiating events rather than baseline narcosis (Klüver et al., 2016). Hence, the compounds mechanistic diversity is not fully captured in the models, which is reflected in higher uncertainty, occasional non-predictions, and underestimation of toxicity for these chemotypes. Nevertheless, the models remain well-suited for hazard assessment of the majority of chemicals in the dataset. Expanding the database to include underrepresented chemotypes would further enhance structural coverage, improve representation of diverse modes of action, and strengthen the applicability and reliability domains, providing a path for future refinement without compromising the current models' utility.

4.4. Relevance for regulatory assessment and alignment with the OECD (Q)SAR assessment framework

From a regulatory perspective, the present models were developed in line with the OECD (Q)SAR model validation principles that support the recent OECD QAF (OECD, 2024). The endpoint is clearly defined as LC50 at 12 h in *X. laevis* embryos under FETAX conditions, the modelling algorithms (k nearest neighbours regression in istKNN and multiple linear regression in QSARINS) are explicitly described, applicability and reliability domains are characterised for both models, and internal and external validation metrics are reported. In addition, mechanistic considerations are discussed for structurally related groups of chemicals. Taken together, these elements address several of the assessment elements included in the QAF Model Checklist (OECD, 2024).

In addition, a full QAF evaluation, including completion of the Prediction and Result Checklists, was not carried out in this methodological study, because the models are not currently embedded in a specific regulatory context. In the future, inclusions of such QSARs in regulatory or weight of evidence assessments for amphibians, the QAF checklists and associated reporting formats (e.g., QSAR Model Reporting Format

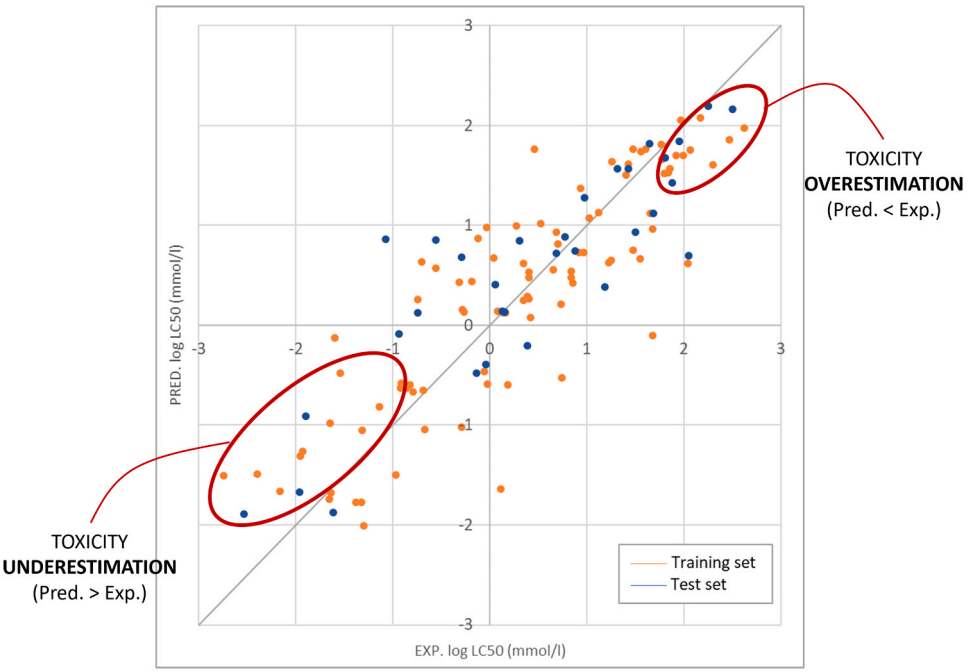

Fig. 10. Superimposed plots describing overlapping experimental vs. predicted training and test sets for acute toxicity in *X. laevis* embryos (LC_{50} 12 h). Training and test sets are represented in orange and blue, respectively. The clusters of datapoints typically subjected to toxicity underestimation (more toxic compounds, on the left) or overestimation (less toxic compounds, on the right) are highlighted in red circles.

and the emerging QSAR Result Reporting Format) should be completed to transparently document the accuracy of inputs, the applicability domain, the reliability of individual predictions and their fitness for purpose (Gissi et al., 2024; OECD, 2024). In this sense, the present work can be considered a first step towards QAF aligned amphibian QSARs, and highlights data and reporting needs that will facilitate their future integration into NAMs and weight of evidence frameworks for ecological risk assessment.

5. Conclusions

The present study addressed data curation of the largest database to date for acute developmental toxicity of chemicals on *Xenopus laevis* embryos and the development and validation of two QSAR models namely k-NN and QSARINS models for the hazard identification and characterisation of chemicals in Xenopus and as a proxy for other amphibian species. Predicting the impact of such environmental chemicals is also of human relevance, according to the “one health” approach since these effects reflect potential neurodevelopmental toxicity particularly in relation to interference with the thyroid axis. In order to build the model, the datasets were curated, and chemical space was explored by means of a principal components analysis using Dragon descriptors. First, a molecular similarity-based approach was taken, by employing a k-NN algorithm.

The k-NN algorithm implemented in istKNN was shown to be a simple method ensuring relatively sound accuracy and prediction performance (R^2 and RMSE of 0.75 and 0.67). Moreover, k-NN results were complemented by further expert mechanistic interpretation using expert knowledge particularly with regard to the scientific basis for under-predictions for highly toxic compounds. Overall, the algorithms were assessed analysed for their main features and limitations including their tendency to assign approximated average values with uncertain predictions. A critical underestimation of the most toxic compounds’ biological activity was highlighted, suggesting the need to increase representation of such molecules. A second method used a Genetic Algorithm for selection of variables in QSARINS. The selected PaDEL descriptors were used to develop the QSAR MLR models, which were then investigated in terms of their Applicability and Reliability Domains. Notably, QSARINS were shown to provide a useful application to perform a range of model validation and provided slightly better predictions compared to those with the k-NN models ((R^2 and RMSE of 0.76 and 0.63). Both models had overall poorer performance with the most toxic compounds, which were underrepresented in the dataset.

Future directions from the present work include the modelling of other endpoints (e.g., EC_{50}, associated with teratogenicity) and the extension of this analysis to other data sources (e.g., other amphibian species) to further strengthen the predictiveness of the model. Moreover, it has been discussed how integration of more highly-toxic compounds may benefit the chemical space coverage and overall performance. Eventually, the database may be expanded to include other developmental stages rather than just the embryonic one. This has been highlighted in a previous work describing the development of a QSAR model for *R. japonica* which proposed the consideration of a taxonomic framework for “true frogs” to identify species as representative of the whole genus and sub-genus for different geographical locations and the relevance of *R. japonica* for the Eurasian species of the subgenus *Rana* "brown frogs" as well as the north American species such as the Northern leopard frog (*Rana pipiens*) and the American bullfrog (*Lithobates catesbeianus*) within their aquatic and terrestrial phase (egg, embryo, tadpole, juvenile and adult). The paper further highlighted that major data gaps in amphibians included lack of chronic toxicity in anuran amphibians and complete lack of toxicity data in other amphibian orders such as Caudata (salamanders and newts) and Gymnophiona (caecilians and relatives). Finally, the limited kinetic information in amphibians was also highlighted for all amphibian taxa particularly to investigate persistence of chemicals. A proposed option to fill in such data gaps was to use fish data as a proxy for amphibians to perform cross-species read-across as a first step. As more data are generated and integrated with quantitative physiological and life cycle data, physiologically-based kinetic models can be developed to estimate bioactive concentrations in amphibians for acute and chronic toxicity and these can also provide opportunities to calibrate, validate and apply dynamic energy budget models for hazard assessment of chemicals at both individual and population level (Baas et al., 2018; Grech et al., 2017; Toropov et al., 2022).

Overall, further research in this field of cheminformatics is promising and it can be foreseen in the near future that the development of further insights on the mechanisms of toxicity and populating public databases will be important to obtain more performant QSAR and in silico models, and ultimately to develop NAMs for ecotoxicity assessment while reducing animal testing.

Disclaimer

The views expressed in this paper are those of the authors only and do not represent the views of the European Food Safety Authority and the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

CRediT authorship contribution statement

Edoardo Carnesecchi: Writing – review & editing. **Nynke Kramer:** Writing – review & editing. **Manuel Ortis:** Writing – review & editing, Data curation. **Edoardo Luca Viganò:** Writing – review & editing. **Erika Colombo:** Writing – review & editing. **Jean Lou CM Dorne:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Alessandra Roncaglioni:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matteo Riccardo Di Nicola:** Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis, Data curation, Conceptualization. **Emilio Benfenati:** Writing – review & editing, Validation, Supervision, Methodology. **Christian Novello:** Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis, Data curation, Conceptualization. **Antony Williams:** Writing – review & editing, Data curation. **Ivano Eberini:** Writing – review & editing. **Anna Melina Steinbach:** Writing – review & editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Acknowledgments

Matteo R. Di Nicola, Nynke Kramer and Edoardo Carnesecchi are thankful to the Lush foundation for supporting their research through the Lush prize 2021 in the area of alternative methods to animal testing.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.toxlet.2025.111813.

Data availability

Main data supporting the findings of this study are included in the article and its Supplementary Material. Additional data and materials are available from the corresponding author upon request.

References

Adhikari, C., Mishra, B.K., 2018. Quantitative structure-activity relationships of aquatic narcosis: a review. CAD 14, 7–28. https://doi.org/10.2174/1573409913666170711130304.

Agrawal, V.K., Chaturvedi, S., Abraham, M.H., Khadikar, P.V., 2003. QSAR study on tadpole narcosis. Bioorg. Med. Chem. 11, 4523–4533. https://doi.org/10.1016/S0968-0896(03)00446-2.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. 46, 175–185. https://doi.org/10.1080/00031305.1992.10475879.

ASTM (American Society for Testing and Materials), 1991. Standard Guide for Conducting the Frog Embryo Teratogenesis Assay-Xenopus (FETAX), in: Annual Book of ASTM Standards.

Astuto, M.C., Di Nicola, M.R., Tarazona, J.V., Rortais, A., Devos, Y., Liem, A.K.D., Kass, G.E.N., Bastaki, M., Schoonjans, R., Maggiore, A., Charles, S., Ratier, A., Lopes, C., Gestin, O., Robinson, T., Williams, A., Kramer, N., Carnesecchi, E., Dorne, J.-L.C.M., 2022. In silico methods for environmental risk assessment: principles, tiered approaches, applications, and future perspectives. In: In Silico Methods for Predicting Drug Toxicity, Methods in Molecular Biology. Springer US, New York, NY, pp. 589–636. https://doi.org/10.1007/978-1-0716-1960-5_23.

Baas, J., Augustine, S., Marques, G.M., Dorne, J.-L., 2018. Dynamic energy budget models in ecological risk assessment: from principles to applications. Sci. Total Environ. 628–629, 249–260. https://doi.org/10.1016/j.scitotenv.2018.02.058.

Ballabio, D., 2015. A MATLAB toolbox for principal component analysis and unsupervised exploration of data structure. Chemom. Intell. Lab. Syst. 149, 1–9.

Benfenati, E., Boriani, E., Craciun, M., Malazizi, L., Neagu, D., Roncaglioni, A., 2007. Databases for pesticide ecotoxicity. in: Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes. Elsevier, pp. 59–81. https://doi.org/10.1016/B978-044452710-3/50004-5.

Berg, N., De Wever, B., Fuchs, H.W., Gaca, M., Krul, C., Roggen, E.L., 2011. Toxicology in the 21st century – working our way towards a visionary reality. Toxicol. Vitr. 25, 874–881. https://doi.org/10.1016/j.tiv.2011.02.008.

Bhattacharya, S., Zhang, Q., Carmichael, P.L., Boekelheide, K., Andersen, M.E., 2011. Toxicity testing in the 21st century: defining new risk assessment approaches based on perturbation of intracellular toxicity pathways. PLoS ONE 6, e20887. https://doi.org/10.1371/journal.pone.0020887.

Blum, M., Ott, T., 2018. *Xenopus*: an undervalued model organism to study and model human genetic disease. Cells Tissues Organs 205, 303–313. https://doi.org/10.1159/000490898.

Campbell Grant, E.H., Miller, D.A.W., Muths, E., 2020. A synthesis of evidence of drivers of amphibian declines. Herpetologica 76, 101. https://doi.org/10.1655/0018-0831-76.2.101.

Cannatella, D.C., De Sá, R.O., 1993. *Xenopus laevis* as a model organism. Syst. Biol. 42, 476–507.

Carnesecchi, E., Toma, C., Roncaglioni, A., Kramer, N., Benfenati, E., Dorne, J.L.C.M., 2020. Integrating QSAR models predicting acute contact toxicity and mode of action profiling in honey bees (A. mellifera): data curation using open source databases, performance testing and validation. Sci. Total Environ. 735, 139243. https://doi.org/10.1016/j.scitotenv.2020.139243.

Carotenuto, R., Tussellino, M., Ronca, R., Benvenuto, G., Fogliano, C., Fusco, S., Netti, P. A., 2022. Toxic effects of SiO2NPs in early embryogenesis of *Xenopus laevis*. Chemosphere 289, 133233. https://doi.org/10.1016/j.chemosphere.2021.133233.

Carr, J.A., Patiño, R., 2011. The hypothalamus–pituitary–thyroid axis in teleosts and amphibians: Endocrine disruption and its consequences to natural populations. Gen. Comp. Endocrinol. 170, 299–312. https://doi.org/10.1016/j.ygcen.2010.06.001.

Casati, S., 2018. Integrated approaches to testing and assessment. Basic Clin. Pharma Tox 123, 51–55. https://doi.org/10.1111/bcpt.13018.

Cattaneo, I., Kalian, A.D., Di Nicola, M.R., Dujardin, B., Levorato, S., Mohimont, L., Nathanail, A.V., Carnessechi, E., Astuto, M.C., Tarazona, J.V., Kass, G.E.N., Liem, A. K.D., Robinson, T., Manini, P., Hogstrand, C., Price, P.S., Dorne, J.L.C.M., 2023. Risk assessment of combined exposure to multiple chemicals at the european food safety authority: principles, guidance documents, applications and future challenges. Toxins 15, 40. https://doi.org/10.3390/toxins15010040.

Como, F., Carnesecchi, E., Volani, S., Dorne, J.L., Richardson, J., Bassan, A., Pavan, M., Benfenati, E., 2017. Predicting acute contact toxicity of pesticides in honeybees (Apis mellifera) through a k-nearest neighbor model. Chemosphere 166, 438–444. https://doi.org/10.1016/j.chemosphere.2016.09.092.

Consonni, V., Ballabio, D., Todeschini, R., 2010. Evaluation of model predictive ability by external validation techniques. J. Chemom. 24, 194–201. https://doi.org/10.1002/cem.1290.

Cronin, M.T.D., Madden, J.C., 2010. In Silico Toxicology-An Introduction. In: Cronin, M., Madden, J. (Eds.), In Silico Toxicology. The Royal Society of Chemistry, pp. 1–10. https://doi.org/10.1039/9781849732093-00001.

Daudin, F.M., 1802. Histoire naturelle des rainettes, des grenouilles et des crapauds: dédiée à B.G.E.L. Lacépède. De l'Imprimerie de Bertrandet, Paris, p. 85. https://doi.org/10.5962/bhl.title.5054.

Dawid, I.B., Sargent, T.D., 1988. *Xenopus laevis* in developmental and molecular biology. Science 240, 1443–1448. https://doi.org/10.1126/science.3287620.

Dawson, D.A., Bantle, J.A., 1987. Development of a reconstituted water medium and preliminary validation of the Frog Embryo Teratogenesis Assay— *Xenopus* (FETAX). J. Appl. Toxicol. 7, 237–244. https://doi.org/10.1002/jat.2550070403.

Dawson, D.A., Schultz, T.W., Hunter, R.S., 1996. Developmental toxicity of carboxylic acids to Xenopus embryos: A quantitative structure-activity relationship and computer-automated structure evaluation. Teratog. Carcinog. Mutagen 16, 109–124.

https://doi.org/10.1002/(SICI)1520-6866(1996)16:2<109::AID-TCM5>3.0.CO;2-M.
Di Nicola, M.R., Cattaneo, I., Nathanail, A.V., Carnesecchi, E., Astuto, M.C., Steinbach, M., Williams, A.J., Charles, S., Gestin, O., Lopes, C., Lamonica, D., Tarazona, J.V., Dorne, J.L.C.M., 2023. The use of new approach methodologies for the environmental risk assessment of food and feed chemicals. Curr. Opin. Environ. Sci. Health 31, 100416. https://doi.org/10.1016/j.coesh.2022.100416.
Dorne, J.L.C.M., Cortiñas-Abrahantes, J., Spyropoulos, F., Darney, K., Lautz, L., Louisse, J., Kass, G.E.N., Carnesecchi, E., Liem, A.K.D., Tarazona, J.V., Billat, P., Beaudoin, R., Zeman, F., Bodin, C., Smith, A., Nathanail, A., Di Nicola, M.R., Kleiner, J., Terron, A., Parra-Morte, J.M., Verloo, D., Robinson, T., 2023. TKPlate 1.0: an open-access platform for toxicokinetic and toxicodynamic modelling of chemicals to implement new approach methodologies in chemical risk assessment. EFS2 21. https://doi.org/10.2903/j.efsa.2023.e211101.
Dorne, J.L.C.M., Richardson, J., Livaniou, A., Carnesecchi, E., Ceriani, L., Baldin, R., Kovarich, S., Pavan, M., Saouter, E., Biganzoli, F., Pasinato, L., Zare Jeddi, M., Robinson, T.P., Kass, G.E.N., Liem, A.K.D., Toropov, A.A., Toropova, A.P., Yang, C., Tarkhov, A., Georgiadis, N., Di Nicola, M.R., Mostrag, A., Verhagen, H., Roncaglioni, A., Benfenati, E., Bassan, A., 2021. EFSA's OpenFoodTox: an open source toxicological database on chemicals in food and feed and its future developments. Environ. Int. 146, 106293. https://doi.org/10.1016/j.envint.2020.106293.
Dresser, T.H., Rivera, E.R., Hoffmann, F.J., Finch, R.A., 1992. Teratogenic assessment of four solvents using the frog embryo teratogenesis assay- *xenopus* (FETAX). J. Appl. Toxicol. 12, 49–56. https://doi.org/10.1002/jat.2550120111.
Edginton, A.N., Sheridan, P.M., Stephenson, G.R., Thompson, D.G., Boermans, H.J., 2004. Comparative effects of pH and Vision® herbicide on two life stages of four anuran amphibian species. Environ. Toxicol. Chem. 23, 815–822. https://doi.org/10.1897/03-115.
EFSA, Panel on Plant Protection Products and their Residues (PPR), Ockleford, C., Adriaanse, P., Berny, P., Brock, T., Duquesne, S., Grilli, S., Hernandez-Jerez, A.F., Bennekou, S.H., Klein, M., Kuhl, T., Laskowski, R., Machera, K., Pelkonen, O., Pieper, S., Stemmer, M., Sundh, I., Teodorovic, I., Tiktak, A., Topping, C.J., Wolterink, G., Aldrich, A., Berg, C., Ortiz-Santaliestra, M., Weir, S., Streissl, F., Smith, R.H., 2018. Scientific Opinion on the state of the science on pesticide risk assessment for amphibians and reptiles. EFS2 16. https://doi.org/10.2903/j.efsa.2018.5125.
Escher, B.I., Hermens, J.L.M., 2002. Modes of action in ecotoxicology: their role in body burdens, species sensitivity, QSARs, and mixture effects. Environ. Sci. Technol. 36, 4201–4217. https://doi.org/10.1021/es015848h.
Falaschi, M., Muraro, M., Gibertini, C., Delle Monache, D., Lo Parrino, E., Faraci, F., Belluardo, F., Di Nicola, M.R., Manenti, R., Ficetola, G.F., 2022. Explaining declines of newt abundance in northern Italy. Freshw. Biol. 67, 1174–1187. https://doi.org/10.1111/fwb.13909.
FAO-OIE-WHO, 2019. Taking a multisectoral, one health approach: a tripartite guide to addressing zoonotic diseases in countries. World Health Organization.
Fisher, M.C., Garner, T.W.J., 2020. Chytrid fungi and global amphibian declines. Nat. Rev. Microbiol 18, 332–343. https://doi.org/10.1038/s41579-020-0335-x.
Fisher, M.C., Pasmans, F., Martel, A., 2021. Virulence and pathogenicity of chytrid fungi causing amphibian extinctions. Annu. Rev. Microbiol. 75, 673–693. https://doi.org/10.1146/annurev-micro-052621-124212.
Floris, M., Manganaro, A., Nicolotti, O., Medda, R., Mangiatordi, G.F., Benfenati, E., 2014. A generalizable definition of chemical similarity for read-across. J. Chemin. 6, 39. https://doi.org/10.1186/s13321-014-0039-1.
Fort, D.J., McLaughlin, D.W., Rogers, R.L., Buzzard, B.O., 2003. Evaluation of the developmental toxicities of ethanol, acetaldehyde, and thioacetamide using FETAX. Drug Chem. Toxicol. 26, 23–34. https://doi.org/10.1081/DCT-120017555.
Fort, D.J., Rogers, R.L., Thomas, J.H., Buzzard, B.O., Noll, A.M., Spaulding, C.D., 2004. Comparative sensitivity of *Xenopus tropicalis* and *Xenopus laevis* as test species for the FETAX model. J. Appl. Toxicol. 24, 443–457. https://doi.org/10.1002/jat.997.
Freda, J., 1991. The effects of aluminum and other metals on amphibians. Environ. Pollut. 71, 305–328. https://doi.org/10.1016/0269-7491(91)90035-U.
Friel, N., Pettitt, A.N., 2011. Classification using distance nearest neighbours. Stat. Comput. 21, 431–437. https://doi.org/10.1007/s11222-010-9179-y.
Gadaleta, D., 2014. A k-NN algorithm for predicting oral sub-chronic toxicity in the rat. ALTEX 423–432. https://doi.org/10.14573/altex.1405091.
Gadaleta, D., Vuković, K., Toma, C., Lavado, G.J., Karmaus, A.L., Mansouri, K., Kleinstreuer, N.C., Benfenati, E., Roncaglioni, A., 2019. SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. J. Chemin. 11, 58. https://doi.org/10.1186/s13321-019-0383-2.
Gissi, A., Tcheremenskaia, O., Bossa, C., Battistelli, C.L., Browne, P., 2024. The OECD (Q)SAR assessment framework: a tool for increasing regulatory uptake of computational approaches. Comput. Toxicol. 31, 100326. https://doi.org/10.1016/j.comtox.2024.100326.
Gosner, K.L., 1960. A simplified table for staging anuran embryos and larvae with notes on identification. Herpetologica 6, 183–190.
Gozalbes, R., Vicente De Julián-Ortiz, J., 2018. Applications of chemoinformatics in predictive toxicology for regulatory purposes, especially in the context of the EU REACH Legislation. Int. J. Quant. Struct. Prop. Relatsh. 3, 1–24. https://doi.org/10.4018/IJQSPR.2018010101.
Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701. https://doi.org/10.1002/qsar.200610151.
Gramatica, P., Cassani, S., Chirico, N., 2014. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 35, 1036–1044. https://doi.org/10.1002/jcc.23576.
Gramatica, P., Chirico, N., Papa, E., Cassani, S., Kovarich, S., 2013. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J. Comput. Chem. 34, 2121–2132. https://doi.org/10.1002/jcc.23361.
Grech, A., Brochot, C., Dorne, J.-L., Quignot, N., Bois, F.Y., Beaudouin, R., 2017. Toxicokinetic models and related tools in environmental risk assessment of chemicals. Sci. Total Environ. 578, 1–15. https://doi.org/10.1016/j.scitotenv.2016.10.146.
Green, D.M., Lannoo, M.J., Lesbarrères, D., Muths, E., 2020. Amphibian population declines: 30 years of progress in confronting a complex problem. Herpetologica 76, 97. https://doi.org/10.1655/0018-0831-76.2.97.
Gross, J.A., Johnson, P.T.J., Prahl, L.K., Karasov, W.H., 2009. Critical period of sensitivity for effects of cadmium on frog growth and development. Enviro Toxic. Chem. 28, 1227–1232. https://doi.org/10.1897/08-205.1.
Hanser, T., Barber, C., Marchaland, J.F., Werner, S., 2016. Applicability domain: towards a more formal definition. SAR QSAR Environ. Res. 27, 865–881. https://doi.org/10.1080/1062936X.2016.1250229.
Huang, H., Wang, X., Ou, W., Zhao, J., Shao, Y., Wang, L., 2003. Acute toxicity of benzene derivatives to the tadpoles (Rana japonica) and QSAR analyses. Chemosphere 53, 963–970. https://doi.org/10.1016/S0045-6535(03)00715-X.
Jaiswal, M., Khadikar, P., 2004. QSAR study on tadpole narcosis using PI index: a case of heterogenous set of compounds. Bioorg. Med. Chem. 12, 1731–1736. https://doi.org/10.1016/j.bmc.2004.01.009.
Klüver, N., Vogs, C., Altenburger, R., Escher, B.I., Scholz, S., 2016. Development of a general baseline toxicity QSAR model for the fish embryo acute toxicity test. Chemosphere 164, 164–173. https://doi.org/10.1016/j.chemosphere.2016.08.079.
Lavado, G.J., Baderna, D., Gadaleta, D., Ultre, M., Roy, K., Benfenati, E., 2021. Ecotoxicological QSAR modeling of the acute toxicity of organic compounds to the freshwater crustacean Thamnocephalus platyurus. Chemosphere 280, 130652. https://doi.org/10.1016/j.chemosphere.2021.130652.
Lushchak, V.I., 2011. Adaptive response to oxidative stress: Bacteria, fungi, plants and animals. Comp. Biochem. Physiol. Part C Toxicol. Pharmacol. 153, 175–190. https://doi.org/10.1016/j.cbpc.2010.10.004.
Manganaro, A., Pizzo, F., Lombardo, A., Pogliaghi, A., Benfenati, E., 2016. Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm. Chemosphere 144, 1624–1630. https://doi.org/10.1016/j.chemosphere.2015.10.054.
National Research Council, 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. National Academies Press, Washington, D.C. https://doi.org/10.17226/11970.
Nikos, P., 2012. *Xenopus laevis* as a model system. Mater Methods 2. https://doi.org/10.13070/mm.en.2.151.
OECD, 2014. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment. OECD. https://doi.org/10.1787/9789264085442-en.
OECD, 2015. Test No. 241: The Larval Amphibian Growth and Development Assay (LAGDA), OECD Guidelines for the Testing of Chemicals, Section 2. OECD. https://doi.org/10.1787/9789264242340-en.
OECD, 2024. (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions, Second Edition, OECD Series on Testing and Assessment, No. 405. OECD. https://doi.org/10.1787/bbdac345-en.
Ortiz-Santaliestra, M.E., Maia, J.P., Egea-Serrano, A., Brühl, C.A., Lopes, I., 2017. Biological relevance of the magnitude of effects (considering mortality, sub-lethal and reproductive effects) observed in studies with amphibians and reptiles in view of population level impacts on amphibians and reptiles. EFS3 14. https://doi.org/10.2903/sp.efsa.2017.EN-1251.
Ortiz-Santaliestra, M.E., Maia, J.P., Egea-Serrano, A., Lopes, I., 2018. Validity of fish, birds and mammals as surrogates for amphibians and reptiles in pesticide toxicity assessment. Ecotoxicology 27, 819–833. https://doi.org/10.1007/s10646-018-1911-y.
Palomar, G., Fernández-Chacón, A., Bosch, J., 2023. Amphibian survival compromised by long-term effects of chytrid fungus. Biodivers. Conserv. 32, 793–809. https://doi.org/10.1007/s10531-022-02525-3.
QSARINS, 2019. v 2.2.4 Manual. ⟨www.qsar.it⟩.
Quigley, I.K., Stubbs, J.L., Kintner, C., 2011. Specification of ion transport cells in the Xenopus larval skin. Development 138, 705–714. https://doi.org/10.1242/dev.055699.
Raies, A.B., Bajic, V.B., 2016. *In silico* toxicology: computational methods for the prediction of chemical toxicity. WIREs Comput. Mol. Sci. 6, 147–172. https://doi.org/10.1002/wcms.1240.
Raunio, H., 2011. In silico toxicology – non-testing methods. Front. Pharm. 2. https://doi.org/10.3389/fphar.2011.00033.
Reed, B.T., 2005. Guidance on the housing and care of the African clawed frog *Xenopus laevis*. Report. Royal Society for the Prevention of Cruelty to Animals, Horsham, United Kingdom.
Rollins-Smith, L.A., 2020. Global amphibian declines, disease, and the ongoing battle between batrachochytrium fungi and the immune system. Herpetologica 76, 178. https://doi.org/10.1655/0018-0831-76.2.178.
Russell, W.M.S., Burch, R.L., 1992. The principles of humane experimental technique, Special ed., [Nachdr. der Ausg.] London 1959. Universities Federation for Animal Welfare, Potters Bar, Herts.
Sahoo, S., Adhikari, C., Kuanar, M., Mishra, B., 2016. A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. CAD 12, 181–205. https://doi.org/10.2174/1573409912666160525112114.

Sparling, D.W., Fellers, G., 2007. Comparative toxicity of chlorpyrifos, diazinon, malathion and their oxon derivatives to larval Rana boylii. Environ. Pollut. 147, 535–539. https://doi.org/10.1016/j.envpol.2006.10.036.

Straka, H., Simmers, J., 2012. *Xenopus laevis*: an ideal experimental model for studying the developmental dynamics of neural network assembly and sensory-motor computations. Dev. Neurobiol. 72, 649–663. https://doi.org/10.1002/dneu.20965.

Tindall, A.J., Morris, I.D., Pownall, M.E., Isaacs, H.V., 2007. Expression of enzymes involved in thyroid hormone metabolism during the early development of *Xenopus tropicalis*. Biol. Cell 99, 151–163. https://doi.org/10.1042/BC20060074.

Toma, C., Cappelli, C.I., Manganaro, A., Lombardo, A., Arning, J., Benfenati, E., 2021. New models to predict the acute and chronic toxicities of representative species of the main trophic levels of aquatic environments. Molecules 26, 6983. https://doi.org/10.3390/molecules26226983.

Toropov, A.A., Di Nicola, M.R., Toropova, A.P., Roncaglioni, A., Carnesecchi, E., Kramer, N.I., Williams, A.J., Ortiz-Santaliestra, M.E., Benfenati, E., Dorne, J.-L.C.M., 2022. A regression-based QSAR-model to predict acute toxicity of aromatic chemicals in tadpoles of the Japanese brown frog (Rana japonica): calibration, validation, and future developments to support risk assessment of chemicals in amphibians. Sci. Total Environ. 830, 154795. https://doi.org/10.1016/j.scitotenv.2022.154795.

Toropov, A.A., Di Nicola, M.R., Toropova, A.P., Roncaglioni, A., Dorne, J.L.C.M., Benfenati, E., 2023. Quasi-SMILES: Self-consistent models for toxicity of organic chemicals to tadpoles. Chemosphere 312, 137224. https://doi.org/10.1016/j.chemosphere.2022.137224.

Toropov, A.A., Toropova, A.P., Benfenati, E., 2020. QSAR model for pesticides toxicity to Rainbow Trout based on “ideal correlations. Aquat. Toxicol. 227, 105589. https://doi.org/10.1016/j.aquatox.2020.105589.

Toropova, A.P., Toropov, A.A., Roncaglioni, A., Benfenati, E., 2021. The index of ideality of correlation improves the predictive potential of models of the antioxidant activity of tripeptides from frog skin (Litoria rubella). Comput. Biol. Med. 133, 104370. https://doi.org/10.1016/j.compbiomed.2021.104370.

Verhaar, H.J.M., Van Leeuwen, C.J., Hermens, J.L.M., 1992. Classifying environmental pollutants. Chemosphere 25, 471–491. https://doi.org/10.1016/0045-6535(92)90280-5.

Wang, L., Xing, P., Wang, C., Zhou, X., Dai, Z., Bai, L., 2018. Maximal information coefficient and support vector regression based nonlinear feature selection and QSAR modeling on toxicity of alcohol compounds to tadpoles of rana temporaria. J. Braz. Chem. Soc. https://doi.org/10.21577/0103-5053.20180176.

Wang, S., Yan, L.C., Zheng, S.S., Li, T.T., Fan, L.Y., Huang, T., Li, C., Zhao, Y.H., 2019. Toxicity of some prevalent organic chemicals to tadpoles and comparison with toxicity to fish based on mode of toxic action. Ecotoxicol. Environ. Saf. 167, 138–145. https://doi.org/10.1016/j.ecoenv.2018.09.105.

Wheeler, G.N., Brändli, A.W., 2009. Simple vertebrate models for chemical genetics and drug discovery screens: lessons from zebrafish and *Xenopus*. Dev. Dyn. 238, 1287–1308. https://doi.org/10.1002/dvdy.21967.

Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474. https://doi.org/10.1002/jcc.21707.

Zhang, Y.-F., Xu, H.-M., Yu, F., Yang, H.-Y., Jia, D.-D., Li, P.-F., 2019. Comparison the sensitivity of amphibian metamorphosis assays with NF 48 stage and NF 51 stage *Xenopus laevis* tadpoles. Toxicol. Mech. Methods 29, 421–427. https://doi.org/10.1080/15376516.2019.1579291.