Science of the Total Environment 942 (2024) 173754

Contents lists available at ScienceDirect

Science of the Total Environment

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

Chaotic neural network algorithm with competitive learning integrated with partial Least Square models for the prediction of the toxicity of fragrances in sanitizers and disinfectants

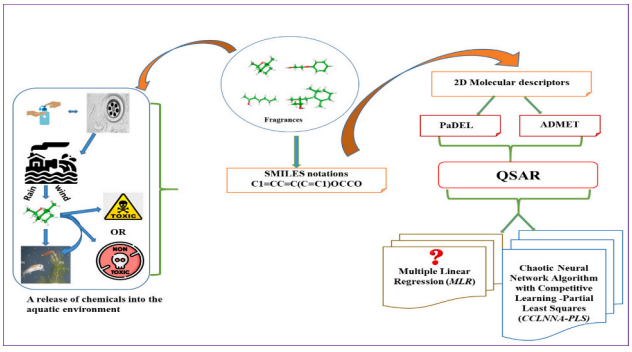
Matshidiso Lephalala [a], Salvador Sagrado Vives [b,c], Krishna Bisetty [a,*]

[a] Department of Chemistry, Durban University of Technology, P.O. Box 1334, Durban 4000, South Africa
[b] Departamento de Química Analítica, Facultad de Farmacia. Universitat de València, E-46100 Burjassot, Valencia, Spain
[c] Instituto Interuniversitario de Investigación de Reconocimiento Molecular y Desarrollo Tecnológico (IDM), Universitat Politècnica de València, Universitat de València, Valencia, Spain

HIGHLIGHTS

- New Chaotic Neural Network Algorithm with Competitive Learning (CCLNNA)-PLS modeling proposed
- Optimized PLS models for fragrance toxicity prediction from structural data
- Comparative stability of PLS models influenced by significant descriptors in optimization
- Enhanced toxicity prediction through dataset expansion to address environmental impact demands

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Philip Hopke

Keywords:
Multiple linear regression (MLR)
Partial least squares (PLS)
Chaotic neural network algorithm with competitive learning (CCLNNA)
Quantitative structure-activity relationship (QSAR)
Toxicity prediction

ABSTRACT

This study addresses the need for accurate structural data regarding the toxicity of fragrances in sanitizers and disinfectants. We compare the predictive and descriptive (model stability) potential of multiple linear regression (MLR) and partial least squares (PLS) models optimized through variable selection (VS). A novel hybrid chaotic neural network algorithm with competitive learning (CCLNNA)–PLS modeling strategy can offer specific optimization with satisfactory results, even for a limited dataset. While also exploring the preliminary comparative analysis, the goal is to introduce an adapted novel CCLNNA optimization strategy for VS, inspired by neural networks, along with exploring the influence of the percentage of significant descriptors in the optimization function to enhance the final model's capabilities.

We analyzed an available dataset of 24 molecules, incorporating ADMET and PaDEL descriptors as predictor variables, to explore the relationship between the response/target variable (pLC_{50}) and the meticulously optimized set of descriptors. The suitability of the selected PLS models (cross- and external-validated accuracy combined with percentage of significant descriptors at a level equal to or >80 %) underscores the importance of

* Corrresponding author.
E-mail address: bisettyk@dut.ac.za (K. Bisetty).

https://doi.org/10.1016/j.scitotenv.2024.173754
Received 13 March 2024; Received in revised form 18 May 2024; Accepted 2 June 2024
Available online 5 June 2024

expanding the dataset to amplify the validation protocols, thus enhancing future model reliability and environmental impact.

1. Introduction

During the COVID-19 pandemic, the World Health Organization (WHO) issued advice on the use of sanitizers and disinfectants as part of non-pharmaceutical interventions (World Health, 2019; Flaxman et al., 2020; Lai et al., 2020; Ngonghala et al., 2020). This led to the use and release of large quantities of chemicals into the aquatic environment, where a portion may have deleterious effects on aquatic taxa (World Health Organization. Regional Office for the Western, 2020; Zamir et al., 2020; Anan et al., 2023). Therefore, concerns have been raised about the potentially deleterious effects of constituent compounds in sanitizers and disinfectants on the aquatic environment (Connon et al., 2012; Guillén et al., 2012; Ebele et al., 2017; Yu, 2020; Musee et al., 2023). Achieving a fine balance between the protection of ecological integrity and the safeguarding of the social and health benefits of these products raises the need to establish their effects and occurrence in the aquatic environment to aid systematic risk assessment.

It is essential to regulate the toxicity of chemicals so that the environment is protected from the harmful effects of chemical exposure. These chemicals can be lethal to species in aquatic ecosystems, which form the food chain; therefore, understanding the toxic effects of chemicals at different levels of the food chain is crucial for risk assessment (McCarty et al., 2018). Due to its increasing consumption, the fathead minnow is a strong bioindicator of human health hazards (Fao., 2016). The Organization for Economic Cooperation and Development (OECD) guideline 203 is used to test for acute toxicity in fish, which involves exposing fish to a chemical for 96 h to determine the LC_{50} or the dose that causes 50 % mortality (Co-operation and Development, 2019). However, toxicology-based experiments are expensive, time-consuming, and laborious. Globally, there is a call to reduce or even ban the use of animal testing for toxicity assessment (de Ávila and Valadares, 2019; Madden et al., 2020; Taylor and Rego Alvarez, 2020). Therefore, in silico techniques have been introduced as alternative approaches because they are inexpensive, robust, and fast (Höfer et al., 2004; Peric et al., 2015; Drgan et al., 2016) and are increasingly being accepted in the regulatory domain (Gellatly and Sewell, 2019; Hardy et al., 2019; Cronin et al., 2022). Among the alternative approaches is the quantitative structure-activity relationship (QSAR) method owing to its inherent advantages (Goudarzi and Goodarzi, 2008; Elmi et al., 2009; Goudarzi et al., 2009; Goudarzi et al., 2012; Satpathy, 2019; Li et al., 2020; Janicka and Śliwińska, 2022; Mozafari et al., 2022a; Mozafari et al., 2022b). The QSAR approach can be used to predict the acute toxicity of these chemicals in fish by establishing a correlation between the molecular structure of these chemicals and their properties (Hall et al., 1989; Veith et al., 1989; Jaworska et al., 1995; Russom et al., 1997; Bearden and Schultz, 1998; Papa et al., 2005; Furusjö et al., 2006; Toropova et al., 2012; Cherkasov et al., 2014; Gramatica et al., 2014; Cappelli et al., 2015; Drgan et al., 2016; Jia et al., 2018; Chen et al., 2020; Jana et al., 2020; Lunghini et al., 2020; Wang and Chen, 2020; Chen et al., 2023).

QSAR models have been accepted and implemented by global regulatory authorities and agencies such as the US Environmental Protection Agency (EPA) and the European Union (Breton and Boxall, 2003). QSAR models have mostly been applied in drug design applications in recent years; however, they are also being used to screen and prioritize environmental chemicals due to a lack of experimental data in the European REACH chemical regulation (Regulation Evaluation Authorization and Restriction of Chemicals) (Guillén et al., 2012). The OECD validation principles are used for the development of QSAR models to ensure the availability of statistically robust and predictive models applicable to the compounds of interest that were employed to predict the relevant endpoints for risk assessment (OECD, 2004). The objective of this work aligns with the challenge of working with a constrained number of available molecules (24). The focus is on assessing the potential of regression models to establish the relationship between pLC_{50} (response variable) and an optimized selection of predictor variables consisting of ADMET and PaDEL descriptors. Although the objective does not seek to propose a conclusive model, it evaluates the future possibilities of various optimized models based on their predictive and descriptive capacities through a preliminary comparative analysis.

To achieve this objective, the study is designed in two phases. In Phase 1, the predictive and descriptive capabilities of the optimized model (via variable selection, VS) are obtained using leave-one-out cross-validation to ensure robustness without risking representativity issues. Only upon successful completion of Phase 1 (with acceptable metrics) does Phase 2 proceed, in which, while maintaining Phase 1 parameters, the model is validated with five external molecules, calibrating it with the remaining 19 molecules. A satisfactory outcome in both phases is deemed sufficient to address the objective premise: defining the model's potential (anticipating OECD principle 4: appropriate measures of goodness-of-fit, robustness, and predictivity). This approach is based on the scientific hypothesis that a high potential for a given model when based on a limited number of available molecules justifies the effort to advance to the final stage by incorporating more molecules into the study.

To the best of our knowledge, this is the first instance where fragrances incorporated into sanitizers and disinfectants have been predicted, in conjunction with a new hybrid approach: 'CCLNNA–PLS' modeling. CCLNNA is a novel variant of the recent neural network algorithm (NNA) (Sadollah et al., 2018). NNA, a metaheuristic algorithm inspired by artificial neural networks (ANNs), stands apart from conventional ANNs that are typically used for prediction. It incorporates randomness into the ANN framework to address optimization challenges, benefiting from the unique structure of ANNs, which endows it with robust global search capabilities. In the NNA, the best solution obtained at each iteration, known as the temporal optimal solution, serves as the target data. The objective is to minimize the error between these target data and other predicted pattern solutions, effectively guiding these solutions toward the target. This approach allows for dynamic adjustment of solutions, refining the most accurate results over time. Notably, NNA requires only two parameters, population size (nPop) and maximum number of iterations (MaxIt), making it simpler than most existing metaheuristic methods. However, NNA suffers from a slow convergence rate and can become trapped in local optima for complex, multimodal optimization problems.

CCLNNA was proposed to address these limitations (Zhang, 2021). It introduces a competitive learning mechanism that divides the population into two subpopulations: "excellent" and "common." These subpopulations are then optimized using distinct learning strategies. The "excellent" subpopulation undergoes a transfer operation to enhance the local search ability, while the "common" subpopulation is updated using a hybrid operator combining transfer and bias operators to improve the global search ability. To further increase the chance of escaping local optima, CCLNNA incorporates a classical chaotic map, the logistic map, to adjust the running probability of the different learning strategies.

The original CCLNNA paper compared its performance with that of the NNA and several recent metaheuristic algorithms using 30 numerical optimization problems and three real-world constrained engineering problems. This comparison highlighted the following main advantages: enhanced global search ability, improved convergence performance, ability to escape local optima, and ability to effectively explore the solution space. Additionally, it outperforms other competitive

optimization methods in terms of solution quality, convergence speed, and stability. However, CCLNNA also has limitations: increased complexity compared to simpler optimization methods due to its additional mechanisms (population division, chaos theory) and thus greater computational overhead.

CCLNNA was conceived as an optimization method (not linked to any predictive model). However, its favorable performance compared to that of other recent methods prompted us to explore its application with toxicological data and to optimize a PLS regression model. With the aim of automating the process, we incorporated the selection of validation objects for the PLS (for Phase 2) into the optimization framework. Additionally, given our interest in performing VS on the descriptors, we adapted the algorithm to encompass this task as well.

The tools outlined in this study can improve our limited knowledge about the hazards associated with fragrances incorporated in sanitizers and disinfectants, particularly their potential risks, such as aquatic toxicity and environmental impact (Musee et al., 2023).

2. Materials and methods

2.1. Dataset creation

In accordance with OECD principle 1 (a defined endpoint), the experimental values for the -logLC_{50} (pLC_{50}) were gathered from the Quantitative Structure-Activity Relationship (QSAR) database and are summarized in Table 1 (Bearden and Schultz, 1997; Genoni, 1997; Devillers, 2001; Papa et al., 2005; Raevsky et al., 2008; Api et al., 2019; Api et al., 2020). The 2D structures were obtained from ChemSpider (ChemSpider: The Free Chemical Database, 2012). The simplified molecular input line entry system (SMILES) representations for the 24 fragrances were acquired from PubChem (Weininger, 1988). Additionally, using an online SMILES Translator and Structure File Generator (https://cactus.nci.nih.gov/translate/), these representations were converted into protein data bank (PDB) structures.

2.2. Calculation of molecular descriptors

A molecular descriptor is a meaningful numerical representation of a molecule that results from a logical and mathematical procedure. It is derived from the symbolic representation of a molecule and is used to represent the structural and physicochemical characteristics of a molecule (Fernández-Torras et al., 2022; Singh et al., 2023). The first 20 descriptors were based on the ADMET parameters calculated using Qikprop, a panel of Maestro in Schrodinger (Lanka et al., 2023). The remaining molecular descriptors were found using the SMILES structural features and the PaDEL-Descriptor tool (v 2.21), which is built into the QSARINS software (Yap, 2011). The calculated molecular descriptors underwent refinement, including the removal of constant or highly similar variables and those with correlations >0.95. The final database comprised the response variable (pLC_{50}) and a total of 393 descriptors obtained using QSARINS software. The data matrix was ordered in ascending order of pLC_{50}. This matrix is ready for further VS processing.

2.3. Comparative analysis of the predictive and descriptive capabilities of algorithms

Initially, the potential of a simplified MLR model (the simplest model) was evaluated through VS. MLR models, along with the use of genetic algorithms (GAs) for variable selection and their combination (GA-MLR), were executed using QSARINS software (Gramatica et al., 2013; Jawarkar et al., 2022). Subsequently, the potential of a more complex but equally transparent model, partial least squares (PLS; compatible with OECD Principle 2: an unambiguous algorithm) (Martens and Martens, 2001; Sagrado and Cronin, 2008), was assessed using variable selection via the recently published optimization algorithm chaotic neural network algorithm with competitive learning (CCLNNA) (Zhang, 2021). Both the PLS models and the CCLNNA optimization algorithm, along with their combination (CCLNNA-PLS), were implemented in MATLAB R2022b. The CCLNNA algorithm was adapted from the version available on the MATLAB Central File Exchange on the MathWorks® website (https://www.mathworks.com/MATLABcentral/fileexchange/100601-cclnna-for-global-optimization). The same descriptors were used in both processes employed in the study (GA-MLR and CCLNNA-PLS).

In Phase 1, the statistical metrics employed to measure predictive capability included R^2 (calibrated with all 24 molecules listed in Table 1) and Q^2_{loo} (a version of R^2 using values of pLC_{50} predicted by PLS for individual molecules from submodels calibrated with the remaining 23 molecules).

In Phase 2, maintaining the parameters of the model calibrated with the 24 molecules listed in Table 1, a model calibrated with 19 training molecules was created and used to predict the remaining 5 external validation molecules. The statistics now included R^2_{tr} (a version of R^2 using training data) and R^2_{ext} (a version of R^2 associated with externally estimated pLC_{50} values). In all cases, autoscaled data were used for PLS.

Additionally, statistics related to the descriptive capability of the model, such as k (the optimal number of latent variables, applicable only to PLS), the number of selected variables, and Psig (the percentage of significant variables), were employed. Psig represents the variables whose uncertainty intervals associated with their regression coefficients do not include zero (Sagrado and Cronin, 2008).

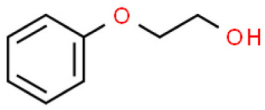
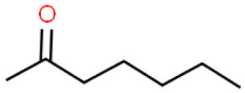
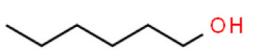
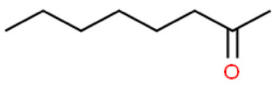
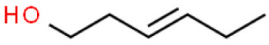
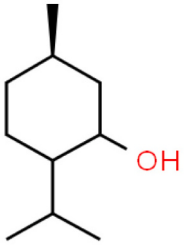
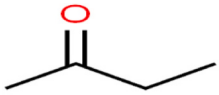
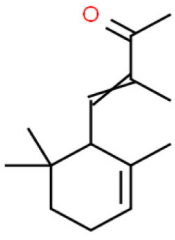
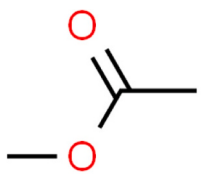
As stated in section 1, CCLNNA (like its predecessor, NNA) requires setting only two main parameters: population size (nPop) and maximum iterations (MaxIt). Prior experiments have shown that a population size (nPop) of 100 and a maximum number of iterations (MaxIt) of 150 are sufficient to achieve satisfactory results. However, in this work, our CCLNNA version (which optimizes a PLS) necessitates additional modifications. Unlike the original algorithm, our adapted CCLNNA transforms the solution vector from continuous variables into discrete integers. For this study, the solution vector comprises 5 indices representing validation molecules (as numbered in Table 1) to be used in Phase 2 and 393 indices for descriptors. This is accomplished by segmenting the indices of molecules into five distinct groups, expanding the pLC_{50} space. Subsequently, the algorithm assigns one index from each group to compose the validation subset. Descriptor indices are converted to binary values (0 or 1) using a cutoff of 0.8, signifying exclusion or inclusion in the model, respectively.

3. Results and discussion

Often, it is challenging to obtain enough experimental data, especially for molecules that have information on the response variable (in this case, pLC_{50}), to construct reliable predictive regression models. However, in such cases, it is possible to assess the potentiality (a priori quality in a preliminary stage) of different models, anticipating the practicality of accumulating enough molecules for a model with guarantees in the final stage.

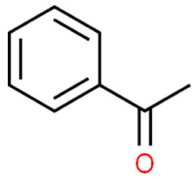
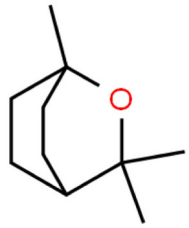
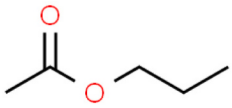
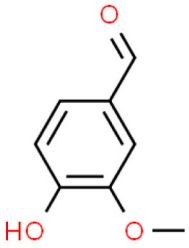
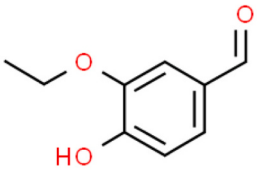
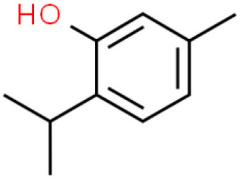
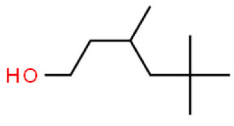
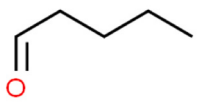
In this work, given the presence of 24 molecules listed in Table 1 with pLC_{50} information, we positioned ourselves exclusively in this preliminary stage. Within this framework, the objective was to evaluate models of increasing complexity until they achieved indicators of sufficiently satisfactory quality, which could activate access to the final stage. For simplicity, MLR models were evaluated first, and more complex or flexible models, such as PLS, were considered only when the quality indicators were unsatisfactory. Only if these models also fail (due to excessive complexity or nonlinearities) would the exploration of predictive models based on artificial neural networks (ANNs) would be considered (although in this case, the transparency of the model would be unclear, unfitting OECD principles).

Table 1
List of fragrances included in the dataset.

No	Name	2D Structure	SMILES	pLC_{50}	References
1	2-Phenoxyethanol		C1 = CC=C(C=C1)OCCO	−3.4	(Raevsky et al., 2008)
2	Heptan-2-one		CCCCCC(=O)C	−3.06	(Raevsky et al., 2008)
3	Hexan-1-ol		CCCCCCO	−2.98	(Raevsky et al., 2008)
4	Octan-2-one		CCCCCCC(=O)C	−2.45	(Raevsky et al., 2008)
5	(*E*)-Hex-3-en-1-ol		CCC=CCCO	−0.58	(Bearden and Schultz, 1997)
6	(1*R*,2*S*,5*R*)-(−)-Menthol		CC1CCC(C(C1)O)C(*C*)C	0.92	(Devillers, 2001)
7	Butan-2-one		CCC(=O)C	1.35	(Papa et al., 2005)
8	3-Methyl-4-(2,6,6-trimethyl-2-cyclohexen-1-yl)-3-buten-2-one		CC1 = CCCC(C1C=C(*C*)C(=O)C)(C)C	1.38	(Api et al., 2019)
9	Methyl acetate		CC(=O)OC	2.32	(Papa et al., 2005)

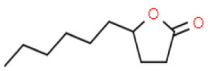
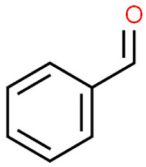
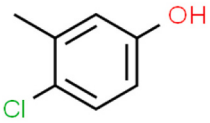
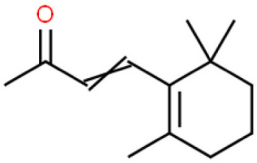
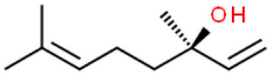
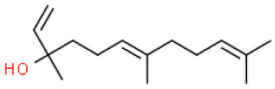
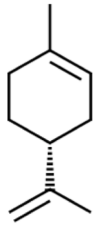
(*continued on next page*)

Table 1 (*continued*)

No	Name	2D Structure	SMILES	pLC_{50}	References
10	1-Phenylethanone		CC(=O)C1 = CC=CC=C1	2.87	(Papa et al., 2005)
11	Eucalyptol		CC1(C2CCC(O1)(CC2)C)C	3.18	(Papa et al., 2005)
12	Propyl acetate		CCCOC(=O)C	3.23	(Papa et al., 2005)
13	Vanillin		COC1 = C(C=CC(=C1)C=O)O	3.26	(Papa et al., 2005)
14	Ethyl Vanillin		CCOC1 = C(C=CC(=C1)C=O)O	3.28	(Papa et al., 2005)
15	Thymol		CC1 = CC(=C(C=C1)C(*C*)C)O	3.30	(Devillers, 2001)
16	3,5,5-Trimethyl-1-hexanol		CC(*CCO*)CC(C)(C)C	3.71	(Genoni, 1997)
17	Pentanal		CCCCC=O	3.83	(Devillers, 2001)

(*continued on next page*)

Table 1 (*continued*)

No	Name	2D Structure	SMILES	pLC_{50}	References
18	Gamma-decalactone		CCCCCCC1CCC(=O)O1	3.98	(Papa et al., 2005)
19	Benzaldehyde		C1 = CC=C(C=C1)C=O	4.03	(Papa et al., 2005)
20	4-Chloro-3-methylphenol		CC1 = C(C=CC(=C1)O)Cl	4.42	(Papa et al., 2005)
21	Beta- Ionone		CC1 = C(C(CCC1)(C)C)C=CC(=O)C	4.58	(Papa et al., 2005)
22	Linalool		CC(=CCCC(C)(C=C)O)C	5.19	(Papa et al., 2005)
23	Nerolidol		CC(=CCCC(=CCCC(C)(C=C)O)C)C	5.19	(Papa et al., 2005)
24	D-Limonene		CC1 = CCC(CC1)C(=C)C	5.29	(Genoni, 1997)

3.1. MLR modeling

In Phase 1, MLR models were evaluated and optimized using the QSARINS software, an excellent platform dedicated to exploring models approaching OECD validation principles (Gramatica et al., 2013). Due to the limited number of molecules and in line with the objective of this work, models were tested using all molecules (without external validation, beyond the cross-validation process). This approach provides robust models with minimal impact on representativeness (by eliminating one molecule in each of the 23 submodels).

Using QSARINS, a VS study of up to nine descriptors (GA-MLR) was conducted and optimized to maximize the Q^2_{loo} statistic. The study suggested that models with 3 selected variables yielded the best results. Thus, the best combination of three descriptors was obtained by testing all possible combinations using the "All Subset" method in QSARINS (Gramatica et al., 2013). The resulting MLR model and statistics are as follows:

$$pLC_{50} = -10(\pm 4) - 0.0004(\pm 0.0002)\ \text{ATS5i} + 0.05(\pm 0.02)\ \text{AATS2v} + 12(\pm 4)\ \text{SpMin6\_Bhs} \quad (1)$$

where the values in parentheses represent the confidence interval of each coefficient as a source of uncertainty. Table 2 presents the statistics for this MLR model in Phase 1. An encouraging result is that the Psig value is 100 %, as deduced from eq. (1). However, it is important to note

Table 2
Selected statistics for the models used.

Selected statistics	MLR model	PLS model-1	PLS model-2a	PLS model-2b
k	–	6	5	5
Selected descriptors	3	16	9	14
Psig.	100	69	100	79
R^2	0.72	0.98	0.90	0.95
Q^2_{loo}	0.62	0.94	0.79	0.85
R^2_{tr}	–	0.97	0.90	0.94
$R^2_{Ext.}$ (Validation)	–	> 0.99	0.95	0.95

Statistics: k: number of latent variables (LVs) in the PLS model; Psig.: percentage of "significant" selected descriptors; R^2: explained y-variance using all (24) molecules to train the model (Phase 1). Q^2_{loo}: Explained y-variance using the leave-one-out cross-validation method (23 submodels, each utilizing 23 training subsets and 1 validation subset of molecules in turn; Phase 1). R^2_{tr}: R^2 for 19 training sets (Phase 2). $R^2_{Ext.}$: R^2 but from the predicted values from the 5-molecule validation subset (Phase 2).

that the uncertainties are of the same order of magnitude as the coefficients. Unfortunately, the values indicate limited predictive capability ($R^2 = 0.72$ and $Q^2_{loo} = 0.62$). In our opinion, the relatively high correlation between the descriptors ATS5i and SpMin6_Bhs (−0.79), as well as other significant correlations (0.43 and 0.36) within the correlation matrix, could influence the GA-MLR process and limit the predictive accuracy of the optimized MLR model. Since PLS is not affected by collinearity, it could accommodate a greater number of variables, which could improve its predictive capacity.

From these results, it was inferred that, with the available data, MLR does not yield a sufficiently satisfactory model. Therefore, increasing the number of molecules probably does not solve this problem. The decision at this point is not to advance to Phase 2 with MLR. Instead, a more flexible model, such as PLS (free from issues related to variable correlation), is tested.

3.2. PLS modeling after CCLNNA optimization (CCLNNA-PLS)

As indicated, owing to its initial effectiveness as a global optimizer, the metaheuristic algorithm NNA has been recently modified to enhance convergence to the optimum and to circumvent local optima. These improvements have culminated in the development of the CCLNNA (Zhang, 2021). To the best of our knowledge, this is the first time that the CCLNNA algorithm has been employed in combination with PLS or for VS tasks.

Following the proposed objective, a PLS model with all available molecules was initially tested, combined with leave-one-out cross-validation (Phase 1). The CCLNNA optimization involved choosing five validation molecules (to be further used in Phase 2, external validation was necessary) and the descriptors to be included in the model. As before, the optimization function was initially designed to maximize Q^2_{loo}. Table 2 (PLS-1 model) presents the resulting statistics. The optimized model employs 6 latent variables (LVs) and a total of 16 original variables. The choice of k = 6 LVs indicates a high complexity in the relationship between the response variable (**y**) and the predictor variables (**X**). Upon examining the internal relationship between the **y**-scores and **X**-scores for the first two LVs (not shown), no evidence of nonlinearity is apparent. However, the substantial data dispersion suggests that increasing the value of k is necessary to enhance the predictive capacity of the model. Additionally, the limitation of the MLR model, which utilizes only three original descriptors, may be attributed to the current selection of 16 descriptors.

The results of the PLS model are shown in Fig. 1. Fig. 1(A) shows the good predictive capacity of the model ($R^2 = 0.98$, $Q^2_{loo} = 0.94$; Table 2). Fig. 1(B) corresponds to Phase 2.

In this case, the PLS model is calibrated using the training subset of 19 molecules (maintaining k = 6 and the 16 descriptors selected in Phase 1), and subsequently, the values of the validation (external) subset (five molecules) are predicted. The figure shows the predictions for the training and validation sets. The validation plot is practically indistinguishable from Fig. 1(A), as is the predictive capacity ($R^2_{tr} = 0.97$, $R^2_{ext} > 0.99$; Table 2). This predictive capacity of the model would likely be viewed favorably by many researchers, judging by the many publications showing regression and prediction statistics, without delving deeper.

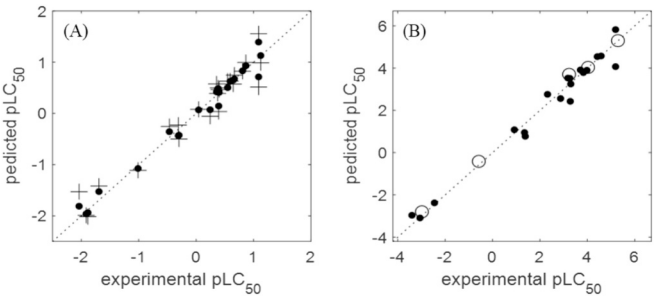


Fig. 1. (A) Validation plots for an optimized CCLNNA-PLS model (PLS model-1) calibrated using all molecules (24), showing the predicted (●) and leave-one-out cross-validated (+) autoscaled values (Phase 1). (B) Predicted values for 19 training molecules (●; used for PLS calibration) and 5 validation molecules (o) used as an external set (Phase 2).

Fig. 2(A) shows the descriptive capability (Martens and Martens, 2001; Sagrado and Cronin, 2008) of PLS model-1, which is based on (i) the values of the regression coefficients and (ii) their uncertainty (**b** ± **U**(**b**)) (Martens and Martens, 2001; Sagrado and Cronin, 2008). Unfortunately, the second aspect, **U**(**b**), which approximates a 95 % confidence interval (Martens and Martens, 2001), is often overlooked. Both are related to the importance of variables in predicting the pLC_{50} value. These descriptive aspects deserve greater attention. Fig. 2(A) shows some small-sized coefficients combined with high uncertainty, rendering them insignificant. Table 2 shows the percentage of significant variables (Psig) calculated for this model (69 %).

This moderate Psig value encourages finding a better balance between predictive and descriptive capabilities. In addition to other possible formulas, a simple method is to introduce Psig into the objective function. For example, Figs. 2(B) and 3 and Table 2 show the results after a CCLNNA-PLS process, generating the PLS-2a model, using the function: 0.8 Q^2_{loo} + 0.2 Psig. This implies that the VS process will tend to sacrifice some predictive capacity to improve descriptive capability (the stability of coefficients). This is not a minor issue, as a more stable model would offer a priori more guarantees when considering the incorporation of more molecules into the study in a further final stage.

The decrease in predictive capacity is noticeable when comparing Figs. 3A and B (PLS model-2a) with Fig. 1 (PLS model-1). A decrease in all the prediction statistics was also observed, especially for Q^2_{loo} (0.79 when it was previously 0.94; Table 2). This raises some doubts about its potential. However, the PLS-2a model operates with a lower k value (5 vs. 6, a simpler model), a smaller number of descriptors (9 versus 16), and, above all, a Psig value of 100 % (more stable).

3.3. Complementary strategies via CCLNNA-PLS

It is important to remember that optimization methods always reach different solutions. For example, the PLS model-2b (a new optimization) was obtained under the same conditions as the previous PLS model-2a. The results are presented in Table 2 and Figs. 2 and 4. In this case, intermediate results were obtained both descriptively and predictively compared to the preceding PLS models, highlighting a Psig value of 79 % (reasonably adequate in this preliminary stage) along with a more encouraging Q^2_{loo} of 0.85 (partially dispelling doubts about its potential) compared to its previous 'replica' of 0.79.

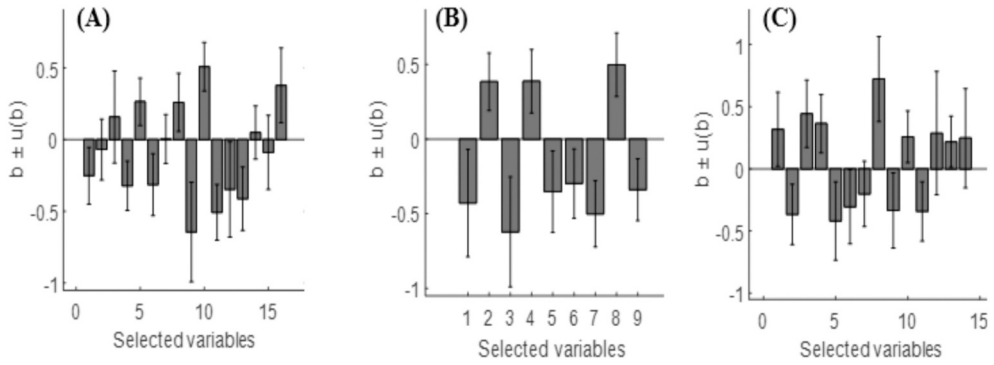


Fig. 2. Importance of the selected variables through their PLS regression coefficients and their corresponding uncertainties from the cross-validation process for (A) PLS model-1, (B) PLS model-2a, and (C) PLS model-2b.

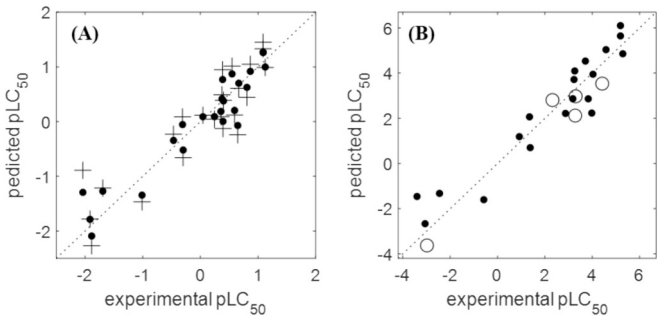


Fig. 3. (A) Validation plots for an optimized CCLNNA-PLS model (PLS model-2a) calibrated using all molecules (24), showing the predicted (●) and leave-one-out cross-validated (+) autoscaled values (Phase 1). (B) Predicted values for 19 training molecules (●; used for PLS calibration) and 5 validation molecules (o) used as an external set (Phase 2).

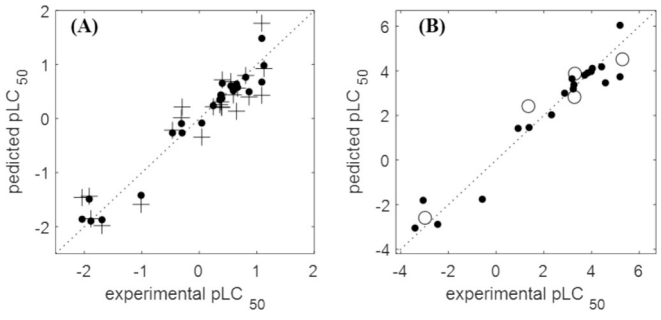


Fig. 4. (A) Validation plots for an optimized CCLNNA-PLS model (PLS model-2b) calibrated using all molecules (24), showing the predicted (●) and leave-one-out cross-validated (+) autoscaled values (Phase 1). (B) Predicted values for 19 training molecules (●; used for PLS calibration) and 5 validation molecules (o) used as an external set (Phase 2).

For the current data and unlike the GA-MLR strategy, this proves that the CCLNNA-PLS strategy is valuable for determining whether there is a good chance of obtaining a model immediately from the start, even with limited data, so that we can decide whether it is worth trying to add new molecules to the model (in the final stage, beyond the current goals). On the other hand, the promising results and the absence of observed nonlinearities suggest that more complex models, such as ANNs (where the descriptive capacity is not yet well resolved), may not be necessary.

Furthermore, in the event of reaching a satisfactory number of molecules, there is the possibility of not settling for a single model but rather accumulating a series of optimized PLS models (which may differ in k, selected descriptors, etc.) and establishing a 'consensus model.' Subsequent predictions with this multiple model set would differ, allowing for the evaluation of the quality of predictions based on result dispersion.

Finally, having this set of models enabled us to examine the frequency of certain descriptors in the set. High frequencies imply that a descriptor is more important than others with null or low frequencies. For example, although outside the objectives of this work (and having obtained very few 'replicas' of the models), Table 3 shows some descriptors that have been repeated along the tested models. This highlights that the PaDEL descriptor SpMin6_Bhs appears in three out of four instances (frequency = 75 %; models MLR, PLS-1, and PLS-2b), which indicates that it has a positive contribution to the prediction of the pLC_{50} values. Within the context of the four example models presented, only one ADMET parameter is used (in the optimized PLS-2b model; it is not used again in the other models). This might suggest that ADMET parameters may not be as useful for predicting pLC_{50} values compared to PaDEL parameters.

3.4. Additional remarks

The OECD guidelines for QSAR modeling indicate the need to characterize the application domain of the model. The software Unscrambler® (The Unscrambler® v.9.7, n.d.) incorporates the utility of the Hotelling T^2 ellipse (a confidence interval around the multivariate Hotelling T^2) on the PLS score plots. The Hotelling T^2 is the multivariate extension of Student's t-distribution and is one of the statistics defined as the applicability domain for cheminformatics models and is also used in multivariate statistical process control (Mehmood, 2016). Using Hotelling T^2 as a criterion for evaluating the applicability domain of PLS models ensures that predictions are reliable and based on data similar to those used to build the model. As an example. Fig. 5 shows the applicability domain of the PLS model-2b (under the conditions depicted in Fig. 4b). The training and external validation molecules are within or brush against the ellipse that defines the applicability domain,

Table 3
Repeated selected descriptors (+) on the models used in Table 1.

Selected descriptors	MLR model	PLS model-1	PLS model-2a	PLS model-2b
SpMin6_Bhs[a]	+	+		+
ATSC1p[b]		+	+	
MATS8c[c]		+		+
VC-5[d]		+		+

[a] Smallest absolute eigenvalue of the burden-modified matrix - n 6/weighted by the relative I-state, PaDEL.
[b] Centered Broto-Moreau autocorrelation - lag 1/weighted by polarizabilities, PaDEL.
[c] Moran autocorrelation - lag 8/weighted by charges, PaDEL.
[d] Valence cluster, order 5, PaDEL.

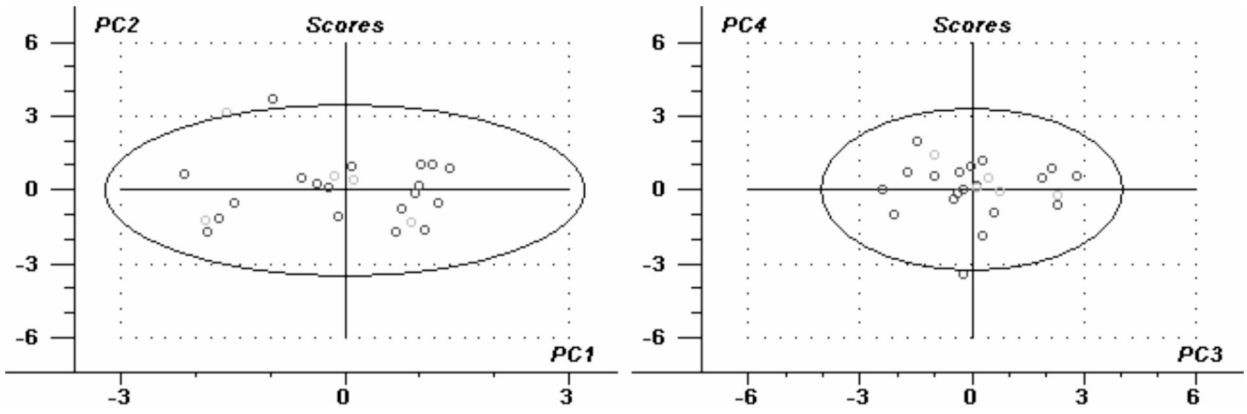


Fig. 5. PLS scores for the four principal latent variables (accounting for 91.3 % of the explained pLC_{50} variance) of PLS model-2b (under the conditions depicted in Fig. 4b), including the Hotelling's T^2 ellipse that defines an applicability domain, as generated by the Unscrambler® software. The validation molecules are represented in light gray.

indicating their reasonable consistency with the model.

The proposed models should be tested to ensure that they are not generated by chance. The so-called "Y-scramble" strategy aims to eliminate chance correlation in the original model. This is achieved by randomly shuffling the responses (pLC_{50} data in this work). The R^2 and Q^2_{loo} values of the model under test must be reasonably higher than those of the "scrambled" models to confirm the absence of chance correlation (Gramatica et al., 2013; Eriksson et al., 2003). In comparison to Table 2, the "scrambled models" (evaluated thrice for each PLS model) yielded inferior statistics; for example, R^2 values varied from 0.06 to 0.64 and Q^2_{loo} values varied from −0.54 to 0.05. Furthermore, the "Y-scramble" strategy adversely affected Psig. values, which ranged from 0 to 21.4. Therefore, we ascertained that the PLS models were not the result of random generation.

Starting from autoscaled descriptor data, the relative importance of descriptors in a model is defined by (i) the absolute value of the regression coefficient, (ii) the sign, related to the direct or inverse contribution in relation to pLC_{50}, and (iii) the relative uncertainty of the coefficient related to its significance (Martens and Martens, 2001; Sagrado and Cronin, 2008). Such information can be obtained from Fig. 2. The overall model significance was assessed by Psig (a discrete statistic). It has proven to be decisive in both the CCLNNA-PLS process, specifically in the VS process, and in the quality of the optimized PLS model. Consequently, the utilization of such statistics or related ones (e.g., continuous statistics centered on the relative uncertainty of the coefficients) should be crucial in future models. Currently, due to the constrained number of available molecules, the coefficients and their uncertainties in Fig. 2 are merely approximations until the future model is finalized.

Mechanistic explanations play a crucial role in QSAR studies by providing an understanding of which chemical compounds exert their toxic effects. On the other hand, in this study, the main goal is to guide decision-making by predicting the potential validity of a future definitive model and determining the cost benefit of incorporating additional data. For instance, the descriptors selected (e.g., those in Table 3) are only approximations until the future model is prepared. Future studies will expand the dataset, and mechanistic insights will then be fundamental in emphasizing the toxicological significance of the structural features identified by QSAR models. Although still in its preliminary stage, just as an example, we can make a provisional approximation. According to the Globally Harmonized System (GHS) for chemical classification and labeling, as outlined in CHEMICALS 2002, 2-phenoxyethanol falls into the most toxic category for fathead minnows under acute classification, leading the list in Table 1 in terms of toxicity ($pLC_{50} = -3.4$). Among the descriptors presented in Table 3, 2-phenoxyethanol has a relatively low SpMin6_Bhs value (0.1) compared to other compounds in Table 1 (SpMin6_Bhs in the 0.07–1.0 range). As indicated, SpMin6_Bhs was the most frequent descriptor in Table 3. As well, a moderate inverse relationship between pLC_{50} and SpMin6_Bhs values is observed. This is compatible with a larger surface area of 2-phenoxyethanol for interacting with lipids in biological membranes (Burka et al., 1997; Zahl et al., 2012), contributing to its toxicity in fathead minnows. This is consistent with the presence of a phenyl group in the 2-phenoxyethanol molecule, which contributes to its lipophilicity. Structurally, the phenyl group contributes to the aromaticity of 2-phenoxyethanol, influencing the arrangement of valence electrons. Furthermore, the ether linkage increases the chemical's water solubility, which may impact the arrangement of valence electrons around the oxygen atom in the ether linkage.

4. Conclusions

This study addresses the challenge of predicting pLC_{50} values from a limited dataset of 24 molecules, emphasizing effective model selection and optimization in the face of data scarcity. The introduction of innovative methodologies, such as using an adapted chaotic neural network algorithm with competitive learning (CCLNNA) along with partial least squares (PLS) modeling, has made the search for pLC_{50} predictions more significant. The primary objective was to predict the potential validity of a conclusive model, guiding decisions on incorporating additional data for a cost benefit analysis in decision-making.

The strategic process outlined in this work, starting with simpler models such as MLR and escalating to more complex PLS models, exemplifies a thoughtful approach tailored to the dataset's constraints. The incorporation of the hybrid CCLNNA-PLS strategy, which emphasizes a balanced consideration of predictive and descriptive aspects, showcases a pioneering approach that could inform future studies in the field of quantitative structure-activity relationship (QSAR) modeling. This research sets the stage for future endeavors seeking to bridge the gap between data limitations and robust predictive modeling in the realm of chemical informatics.

It is important that the PLS models that are made, especially when Psig is involved in the CCLNNA process (changing the optimization function), can accurately predict and remain stable in the chosen descriptors. This implies that expanding the size of the dataset is a realistic strategy.

CRediT authorship contribution statement

Matshidiso Lephalala: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Salvador Sagrado Vives:** Writing – original draft, Validation, Methodology, Formal analysis.

Krishna Bisetty: Writing – review & editing, Supervision, Resources, Project administration.

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.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank the South African Centre for High-Performance Computing (CHPC) for providing the software resources, and the authors are also grateful to Prof. Paola Gramatica from the University of Insubria in Varese, Italy, for providing the QSARINS software used for this study. We gratefully acknowledge the Doctoral Scholarship Support from the Water Research Commission (WRC) in Pretoria and the Durban University of Technology (DUT). ML wishes to express her gratitude to Prof. Ndeke Musee for his assistance in conceptualizing this project.

References

Anan, T., Ishimaru, T., Hino, A., Nagata, T., Tateishi, S., Tsuji, M., Ogami, A., Matsuda, S., Fujino, Y., 2023. Association between COVID-19 infection rates by region and implementation of non-pharmaceutical interventions: a cross-sectional study in Japan. J. Public Health (Oxf.) 45 (1), 229–236. https://doi.org/10.1093/pubmed/fdab385.

Api, A.M., Belsito, D., Biserta, S., Botelho, D., Bruze, M., Burton, G.A., Buschmann, J., Cancellieri, M.A., Dagli, M.L., Date, M., Dekant, W., Deodhar, C., Fryer, A.D., Gadhia, S., Jones, L., Joshi, K., Lapczynski, A., Lavelle, M., Liebler, D.C., Na, M., O'Brien, D., Patel, A., Penning, T.M., Ritacco, G., Rodriguez-Ropero, F., Romine, J., Sadekar, N., Salvito, D., Schultz, T.W., Siddiqi, F., Sipes, I.G., Sullivan, G., Thakkar, Y., Tokura, Y., Tsang, S., 2019. RIFM fragrance ingredient safety assessment, methyl ionone (mixture of isomers), CAS registry number 1335-46-2. Food Chem. Toxicol. 134, 110716 https://doi.org/10.1016/j.fct.2019.110716.

Api, A.M., Belsito, D., Botelho, D., Bruze, M., Burton, G.A., Buschmann, J., Dagli, M.L., Date, M., Dekant, W., Deodhar, C., Francis, M., Fryer, A.D., Jones, L., Joshi, K., La Cava, S., Lapczynski, A., Liebler, D.C., O'Brien, D., Patel, A., Penning, T.M., Ritacco, G., Romine, J., Sadekar, N., Salvito, D., Schultz, T.W., Sipes, I.G., Sullivan, G., Thakkar, Y., Tokura, Y., Tsang, S., 2020. RIFM fragrance ingredient safety assessment, 3-(4-methyl-3-pentenyl)-3-cyclohexene-1-carbonitrile, CAS registry number 68084-04-8. Food Chem. Toxicol. 144, 111491 https://doi.org/10.1016/j.fct.2020.111491.

Bearden, A.P., Schultz, T.W., 1997. Structure-activity relationships for Pimephales and Tetrahymena: A mechanism of action approach. Environ. Toxicol. Chem. 16 (6), 1311–1317. https://doi.org/10.1002/etc.5620160629.

Bearden, A.P., Schultz, T.W., 1998. Comparison of Tetrahymena and Pimephales toxicity based on mechanism of action. SAR QSAR Environ. Res. 9 (3–4), 127–153. https://doi.org/10.1080/10629369808039153.

Breton, R., Boxall, A., 2003. Pharmaceuticals and personal care products in the environment: regulatory drivers and research needs. QSAR Comb. Sci. 22 (3), 399–409. https://doi.org/10.1002/qsar.200390030.

Burka, J.F., Hammell, K.L., Horsberg, T.E., Johnson, G.R., Rainnie, D.J., Speare, D.J., 1997. Drugs in salmonid aquaculture – A review. J. Vet. Pharmacol. Ther. 20 (5), 333–349. https://doi.org/10.1046/j.1365-2885.1997.00094.x.

Cappelli, C.I., Cassano, A., Golbamaki, A., Moggio, Y., Lombardo, A., Colafranceschi, M., Benfenati, E., 2015. Assessment of in silico models for acute aquatic toxicity towards fish under REACH regulation. SAR QSAR Environ. Res. 26 (12), 977–999. https://doi.org/10.1080/1062936X.2015.1104519.

CHEMICALS, L. O, 2002. Globally harmonized system of classification and labeling of chemicals (GHS). https://unece.org/sites/default/files/datastore/fileadmin/DAM/trans/doc/2002/ac10c4/ST-SG-AC10-C4-2002-04e.pdf.

ChemSpider, 2012. The free chemical database. Ref. Rev. 26 (7), 45–46. https://doi.org/10.1108/09504121211271059.

Chen, X., Dang, L., Yang, H., Huang, X., Yu, X., 2020. Machine learning-based prediction of toxicity of organic compounds towards fathead minnow. RSC Adv. 10 (59), 36174–36180. https://doi.org/10.1039/D0RA05906D.

Chen, S., Sun, G., Fan, T., Li, F., Xu, Y., Zhang, N., Zhao, L., Zhong, R., 2023. Ecotoxicological QSAR study of fused/nonfused polycyclic aromatic hydrocarbons (FNFPAHs): assessment and priority ranking of the acute toxicity to *Pimephales promelas* by QSAR and consensus modelling methods. Sci. Total Environ. 876, 162736 https://doi.org/10.1016/j.scitotenv.2023.162736.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A. and Tropsha, A., 2014. QSAR modelling: where have you been? Where are you going to? J. Med. Chem. 57 (12), 4977–5010. https://doi.org/10.1021/jm4004285.

Connon, R.E., Geist, J., Werner, I., 2012. Effect-based tools for monitoring and predicting the Ecotoxicological effects of Chemicals in the Aquatic Environment. Sensors 12 (9), 12741–12771. https://doi.org/10.3390/s120912741.

Co-operation, O. f. E. and Development, 2019. Test No. 203: Fish, Acute Toxicity Test. OECD Publishing. https://doi.org/10.1787/9789264069961-en, 2:2.

Cronin, M.T.D., Enoch, S.J., Madden, J.C., Rathman, J.F., Richarz, A.-N., Yang, C., 2022. A review of in silico toxicology approaches to support the safety assessment of cosmetics-related materials. Computational Toxicology 21, 100213. https://doi.org/10.1016/j.comtox.2022.100213.

de Ávila, R.I., Valadares, M.C., 2019. Brazil moves toward the replacement of animal experimentation. Altern. Lab. Anim. 47 (2), 71–81. https://doi.org/10.1177/0261192919856806.

Devillers, J., 2001. QSAR modelling of large heterogeneous sets of molecules. SAR QSAR Environ. Res. 12 (6), 515–528. https://doi.org/10.1080/10629360108039832.

Drgan, V., Župerl, Š., Vračko, M., Como, F., Novič, M., 2016. Robust modelling of acute toxicity towards fathead minnow (*Pimephales promelas*) using counterpropagation artificial neural networks and genetic algorithm. SAR QSAR Environ. Res. 27 (7), 501–519. https://doi.org/10.1080/1062936X.2016.1196388.

Ebele, A.J., Abou-Elwafa Abdallah, M., Harrad, S., 2017. Pharmaceuticals and personal care products (PPCPs) in the freshwater aquatic environment. Emerging Contaminants 3 (1), 1–16. https://doi.org/10.1016/j.emcon.2016.12.004.

Elmi, Z., Faez, K., Goodarzi, M., Goudarzi, N., 2009. Feature selection method based on fuzzy entropy for regression in QSAR studies. Molecular Physics 107 (17), 1787–1798. https://doi.org/10.1080/00268970903078559.

Eriksson, L., Jaworska, J., Worth Andrew, P., Cronin Mark, T.D., McDowell Robert, M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Perspect. 111 (10), 1361–1375. https://doi.org/10.1289/ehp.5758.

Fao., 2016. State of world fisheries and aquaculture 2016 (russian). Food & Agriculture Org. https://doi.org/10.1088/1755-1315/715/1/012052.

Fernández-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M., Aloy, P., 2022. Connecting chemistry and biology through molecular descriptors. Curr. Opin. Chem. Biol. 66, 102090 https://doi.org/10.1016/j.cbpa.2021.09.001.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J.W., Monod, M., Perez-Guzman, P.N., Schmit, N., Cilloni, L., Ainslie, K.E.C., Baguelin, M., Boonyasiri, A., Boyd, O., Cattarino, L., Cooper, L.V., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Djaafara, B., Dorigatti, I., van Elsland, S.L., FitzJohn, R.G., Gaythorpe, K.A.M., Geidelberg, L., Grassly, N.C., Green, W.D., Hallett, T., Hamlet, A., Hinsley, W., Jeffrey, B., Knock, E., Laydon, D.J., Nedjati-Gilani, G., Nouvellet, P., Parag, K.V., Siveroni, I., Thompson, H.A., Verity, R., Volz, E., Walters, C.E., Wang, H., Wang, Y., Watson, O.J., Winskill, P., Xi, X., Walker, P.G.T., Ghani, A.C., Donnelly, C.A., Riley, S., Vollmer, M.A.C., Ferguson, N.M., Okell, L.C., Bhatt, S., Imperial College, C.-R.T., 2020. Estimating the effects of nonpharmaceutical interventions on COVID-19 in Europe. Nature 584 (7820), 257–261. https://doi.org/10.1038/s41586-020-2405-7.

Furusjö, E., Svenson, A., Rahmberg, M., Andersson, M., 2006. The importance of outlier detection and training set selection for reliable environmental QSAR predictions. Chemosphere 63 (1), 99–108. https://doi.org/10.1016/j.chemosphere.2005.07.002.

Gellatly, N., Sewell, F., 2019. Regulatory acceptance of in silico approaches for the safety assessment of cosmetic-related substances. Computational Toxicology 11, 82–89. https://doi.org/10.1016/j.comtox.2019.03.003.

Genoni, G.P., 1997. Influence of the energy relationships of organic compounds on toxicity to the CladoceranDaphnia magnaand the FishPimephales promelas. Ecotoxicol. Environ. Saf. 36 (1), 27–37. https://doi.org/10.1006/eesa.1996.1481.

Goudarzi, N., Goodarzi, M., 2008. Prediction of the logarithmic partition coefficients (log P) of some organic compounds by least square-support vector machine (LS-SVM). Mol. Phys. 106 (21−23), 2525–2535. https://doi.org/10.1080/00268970802577834.

Goudarzi, N., Goodarzi, M., Hosseini, M.M., Nekooei, M., 2009. QSPR models for prediction of half-wave potentials of some chlorinated organic compounds using SR-PLS and GA-PLS methods. Molecular Physics 107 (17), 1739–1744. https://doi.org/10.1080/00268970903042266.

Goudarzi, N., Goodarzi, M., Chen, T., 2012. QSAR prediction of HIV inhibition activity of styryl quinoline derivatives by genetic algorithm coupled with multiple linear regressions. Med. Chem. Res. 21 (4), 437–443. https://doi.org/10.1007/s00044-010-9542-8.

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 (24), 2121–2132. https://doi.org/10.1002/jcc.23361.

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 (13), 1036–1044. https://doi.org/10.1002/jcc.23576.

Guillén, D., Ginebreda, A., Farré, M., Darbra, R.M., Petrovic, M., Gros, M., Barceló, D., 2012. Prioritization of chemicals in the aquatic environment based on risk assessment: analytical, modeling and regulatory perspective. Sci. Total Environ. 440, 236–252. https://doi.org/10.1016/j.scitotenv.2012.06.064.

Hall, L.H., Maynard, E.L., Kier, L.B., 1989. Structure-activity relationship studies on the toxicity of benzene derivatives: III. Predictions and extension to new substituents. Environ. Toxicol. Chem. 8 (5), 431–436. https://doi.org/10.1002/etc.5620080509.

Hardy, B., Bachler, D., Dokler, J., Exner, T., Hardy, C., Tong, W., Burgwinkel, D., Bergström, R., 2019. OpenTox principles and best practices for trusted reproducible in silico methods supporting research and regulatory applications in toxicological science. In: Hong, H. (Ed.), Advances in Computational Toxicology: Methodologies and Applications in Regulatory Science. Springer International Publishing, Cham, pp. 383–403. Available: https://doi.org/10.1007/978-3-030-16443-0_19.

Höfer, T., Gerner, I., Gundert-Remy, U., Liebsch, M., Schulte, A., Spielmann, H., Vogel, R., Wettig, K., 2004. Animal testing and alternative approaches for the human health risk assessment under the proposed new European chemicals regulation. Arch. Toxicol. 78 (10), 549–564. https://doi.org/10.1007/s00204-004-0577-9.

Jana, G., Pal, R., Sural, S., Chattaraj, P.K., 2020. Quantitative structure-toxicity relationship models based on hydrophobicity and Electrophilicity. In: Roy, K. (Ed.), Ecotoxicological QSARs. Springer US, New York, NY, pp. 661–679. Available: https://doi.org/10.1007/978-1-0716-0150-1_27.

Janicka, M., Śliwińska, A., 2022. Quantitative retention (structure)– activity relationships in predicting the pharmaceutical and toxic properties of potential pesticides. Molecules 27 (11), 3599. https://doi.org/10.3390/molecules27113599.

Jawarkar, R.D., Bakal, R.L., Mukherjee, N., Ghosh, A., Zaki, M.E.A., Al-Hussain, S.A., Al-Mutairi, A.A., Samad, A., Gandhi, A., Masand, V.H., 2022. QSAR evaluations to unravel the structural features in lysine-specific histone demethylase 1A inhibitors for novel anticancer Lead development supported by molecular docking, MD simulation and MMGBSA. Molecules 27 (15). https://doi.org/10.3390/molecules27154758.

Jaworska, J.S., Hunter, R.S., Schultz, T.W., 1995. Quantitative structure-toxicity relationships and volume fraction analyses for selected esters. Arch. Environ. Contam. Toxicol. 29 (1), 86–93. https://doi.org/10.1007/BF00213091.

Jia, Q., Zhao, Y., Yan, F., Wang, Q., 2018. QSAR model for predicting the toxicity of organic compounds to fathead minnow. Environ. Sci. Pollut. Res. 25 (35), 35420–35428. https://doi.org/10.1007/s11356-018-3434-8.

Lai, S., Ruktanonchai, N.W., Zhou, L., Prosper, O., Luo, W., Floyd, J.R., Wesolowski, A., Santillana, M., Zhang, C., Du, X., Yu, H., Tatem, A.J., 2020. Effect of nonpharmaceutical interventions to contain COVID-19 in China. Nature 585 (7825), 410–413. https://doi.org/10.1038/s41586-020-2293-x.

Lanka, G., Bathula, R., Ghosh, B., Potlapally, S.R., 2023. Identification of potential antiviral lead inhibitors against SARS-CoV-2 main protease: structure-guided virtual screening, docking, ADME, and MD simulation-based approach. Artificial Intelligence Chemistry 1 (2), 100015. https://doi.org/10.1016/j.aichem.2023.100015.

Li, X., Kleinstreuer, N.C., Fourches, D., 2020. Hierarchical quantitative structure-activity relationship modelling approach for integrating binary, multiclass, and regression models of acute Oral systemic toxicity. Chem. Res. Toxicol. 33 (2), 353–366. https://doi.org/10.1021/acs.chemrestox.9b00259.

Lunghini, F., Marcou, G., Azam, P., Enrici, M.H., Van Miert, E., Varnek, A., 2020. Consensus QSAR models estimating acute toxicity to aquatic organisms from different trophic levels: algae, Daphnia, and fish. SAR QSAR Environ. Res. 31 (9), 655–675. https://doi.org/10.1080/1062936X.2020.1797872.

Madden, J.C., Enoch, S.J., Paini, A., Cronin, M.T.D., 2020. A review of in silico tools as alternatives to animal testing: principles, resources and applications. Alternatives to laboratory animals: ATLA 48 (4), 146–172. https://doi.org/10.1177/0261192920965977.

Martens, H., Martens, M., 2001. Multivariate analysis of quality. An Introduction. Measurement Science and Technology 12 (10), 1746. https://doi.org/10.1088/0957-0233/12/10/708.

McCarty, L.S., Borgert, C.J., Posthuma, L., 2018. The regulatory challenge of chemicals in the environment: toxicity testing, risk assessment, and decision-making models. Regul. Toxicol. Pharmacol. 99, 289–295. https://doi.org/10.1016/j.yrtph.2018.10.001.

Mehmood, T., 2016. Hotelling T2 based variable selection in partial least squares regression. Chemom. Intel. Lab. Syst. 154, 23–28. https://doi.org/10.1016/j.chemolab.2016.03.001.

Mozafari, Z., Arab Chamjangali, M., Arashi, M., Goudarzi, N., 2022a. QSRR models for predicting the retention indices of VOCs in different datasets using an efficient variable selection method coupled with artificial neural network modeling: ANN-based QSPR modeling. J. Iran. Chem. Soc. 19 (6), 2617–2630. https://doi.org/10.1007/s13738-021-02488-2.

Mozafari, Z., Chamjangali, M.A., Arashi, M., Goudarzi, N., 2022b. Application of the LAD-LASSO as a dimensional reduction technique in the ANN-based QSAR study: discovery of potent inhibitors using molecular docking simulation. Chemom. Intel. Lab. Syst. 222, 104510 https://doi.org/10.1016/j.chemolab.2022.104510.

Musee, N., Ngwenya, P., Motaung, L.K., Moshuhla, K., Nomngongo, P., 2023. Occurrence, effects, and ecological risks of chemicals in sanitizers and disinfectants: A review. Environmental Chemistry and Ecotoxicology 5, 62–78. https://doi.org/10.1016/j.enceco.2023.01.003.

Ngonghala, C.N., Iboi, E., Eikenberry, S., Scotch, M., MacIntyre, C.R., Bonds, M.H., Gumel, A.B., 2020. Mathematical assessment of the impact of nonpharmaceutical interventions on curtailing the 2019 novel coronavirus. Math. Biosci. 325, 108364 https://doi.org/10.1016/j.mbs.2020.108364.

OECD, O. J. C. y. A, 2004. The OECD principles of corporate governance, p. 216. https://doi.org/10.22201/fca.24488410e.2005.562.

Papa, E., Villa, F., Gramatica, P., 2005. Statistically validated QSARs, based on theoretical descriptors, for modelling aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). Journal of Chemical Information and Modelling 45 (5), 1256–1266. https://doi.org/10.1021/ci050212l.

Peric, B., Sierra, J., Martí, E., Cruañas, R., Garau, M.A., 2015. Quantitative structure–activity relationship (QSAR) prediction of (eco)toxicity of short aliphatic protic ionic liquids. Ecotoxicol. Environ. Saf. 115, 257–262. https://doi.org/10.1016/j.ecoenv.2015.02.027.

Raevsky, O.A., Grigor'ev, V.Y., Weber, E.E., Dearden, J.C., 2008. Classification and quantification of the toxicity of chemicals to guppy, fathead minnow and rainbow trout: part 1. Nonpolar narcosis mode of action. QSAR & Combinatorial Science 27 (11−12), 1274–1281. https://doi.org/10.1002/qsar.200860014.

Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E., Drummond, R.A., 1997. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). Environ. Toxicol. Chem. 16 (5), 948–967. https://doi.org/10.1002/etc.5620160514.

Sadollah, A., Sayyaadi, H., Yadav, A., 2018. A dynamic metaheuristic optimization model inspired by biological nervous systems: neural network algorithm. Appl. Soft Comput. 71, 747–782. https://doi.org/10.1016/j.asoc.2018.07.039.

Sagrado, S., Cronin, M.T.D., 2008. Application of the modelling power approach to variable subset selection for GA-PLS QSAR models. Anal. Chim. Acta 609 (2), 169–174. https://doi.org/10.1016/j.aca.2008.01.013.

Satpathy, R., 2019. Quantitative structure–activity relationship methods for the prediction of the toxicity of pollutants. Environ. Chem. Lett. 17 (1), 123–128. https://doi.org/10.1007/s10311-018-0780-1.

Singh, R., Kumar, P., Sindhu, J., Devi, M., Kumar, A., Lal, S., Singh, D., 2023. Parsing structural fragments of thiazolidin-4-one based α-amylase inhibitors: A combined approach employing in vitro colorimetric screening and GA-MLR based QSAR modelling supported by molecular docking, molecular dynamics simulation and ADMET studies. Comput. Biol. Med. 157, 106776 https://doi.org/10.1016/j.compbiomed.2023.106776.

Taylor, K., Rego Alvarez, L., 2020. Regulatory drivers in the last 20 years towards the use of in silico techniques as replacements to animal testing for cosmetic-related substances. Computational Toxicology 13, 100112. https://doi.org/10.1016/j.comtox.2019.100112.

The Unscrambler® v.9.7. 1986-20027 CAMO Software AS. https://www.aspentech.com/en/products/apm/aspen-unscrambler.

Toropova, A.P., Toropov, A.A., Lombardo, A., Roncaglioni, A., Benfenati, E., Gini, G., 2012. Coral: QSAR models for acute toxicity in fathead minnow (*Pimephales promelas*). J. Comput. Chem. 33 (12), 1218–1223. https://doi.org/10.1002/jcc.22953.

Veith, G.D., Lipnick, R.L., Russom, C.L., 1989. The toxicity of Acetylenic alcohols to the fathead minnow, Pimephales Promelas: narcosis and Proelectrophile activation. Xenobiotica 19 (5), 555–565. https://doi.org/10.3109/00498258909042295.

Wang, Y., Chen, X., 2020. A joint optimization QSAR model of fathead minnow acute toxicity based on a radial basis function neural network and its consensus modelling. RSC Adv. 10 (36), 21292–21308. https://doi.org/10.1039/D0RA02701D.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28 (1), 31–36. https://doi.org/10.1021/ci00057a005.

World Health, O, 2019. Non-pharmaceutical Public Health Measures for Mitigating the Risk and Impact of Epidemic and Pandemic Influenza: Annex: Report of Systematic Literature Reviews. World Health Organization, Geneva. Available: https://iris.who.int/handle/10665/329439. (Accessed 18 May 2024).

World Health Organization. Regional Office for the Western, P, 2020. Calibrating Long-Term Non-pharmaceutical Interventions for COVID-19: Principles and Facilitation Tools. WHO Regional Office for the Western Pacific, Manila. Available: http://iris.wpro.who.int/handle/10665.1/14520. (Accessed 18 May 2024).

Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32 (7), 1466–1474. https://doi.org/10.1002/jcc.21707.

Yu, X., 2020. Prediction of chemical toxicity to Tetrahymena pyriformis with four-descriptor models. Ecotoxicol. Environ. Saf. 190, 110146 https://doi.org/10.1016/j.ecoenv.2019.110146.

Zahl, I.H., Samuelsen, O., Kiessling, A., 2012. Anaesthesia of farmed fish: implications for welfare. Fish Physiol. Biochem. 38 (1), 201–218. https://doi.org/10.1007/s10695-011-9565-1.

Zamir, M., Shah, Z., Nadeem, F., Memood, A., Alrabaiah, H., Kumam, P., 2020. Non pharmaceutical interventions for optimal control of COVID-19. Comput. Methods Programs Biomed. 196, 105642 https://doi.org/10.1016/j.cmpb.2020.105642.

Zhang, Y., 2021. Chaotic neural network algorithm with competitive learning for global optimization. Knowl.-Based Syst. 231, 107405 https://doi.org/10.1016/j.knosys.2021.107405.