Chemosphere 312 (2023) 137224

Contents lists available at ScienceDirect

Chemosphere

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

Quasi-SMILES: Self-consistent models for toxicity of organic chemicals to tadpoles

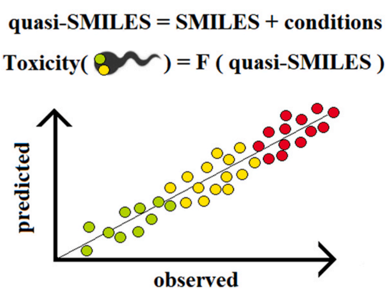
A.A. Toropov [a], M.R. Di Nicola [b], A.P. Toropova [a,*], A. Roncaglioni [a], J.L.C.M. Dorne [c], E. Benfenati [a]

[a] *Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156, Milano, Italy*
[b] *IRCCS San Raffaele Hospital, Unit of Dermatology, Milan, Italy*
[c] *Scientific Committee and Emerging Risks Unit, European Food Safety Authority, Via Carlo Magno 1A, Parma, Italy*

HIGHLIGHTS

- A model of toxicity as a function of all available data is proposed.
- Model calculated with so-called quasi-SMILES.
- Quasi-SMILES is a string encoding observed conditions.
- Conditions are exposure times and kinds of tadpoles.
- The statistical quality of models toxicity to tadpoles is reproducible.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Tamara S. Galloway

Keywords:
Toxicity towards tadpoles
QSAR
Quasi-SMILES
Monte Carlo method
System of self-consistent models

ABSTRACT

Simplified molecular input-line entry systems (SMILES) are the representation of the molecular structure that can be used to establish quantitative structure-property/activity relationships (QSPRs/QSARs) for various endpoints expressed as mathematical functions of the molecular architecture. Quasi-SMILES is extending the traditional SMILES by means of additional symbols that reflect experimental conditions. Using the quasi-SMILES models of toxicity to tadpoles gives the possibility to build up models by taking into account the time of exposure. Toxic effects of experimental situations expressed via 188 quasi-SMILES (the negative logarithm of molar concentrations which lead to lethal 50% tadpoles effected during 12 h, 24 h, 48 h, 72 h, and 96 h) were modelled with good results (the average determination coefficient for the validation sets is about 0.97). In this way, we developed new models for this amphibian endpoint, which is poorly studied.

1. Introduction

Quantitative structure-property/activity relationships (QSPRs/QSARs) gradually become a test-centre for many very distant ideas and aspirations coming from mathematics or from other sciences such as physics, chemistry, biology, and medicine. In mathematics, perturbation theory comprises methods for finding an approximate solution to a problem, and these approaches are very nice tools for solving QSPR/QSAR tasks (González-Díaz et al., 2013). Multi-targets QSARs are an emergence of a new quality in the theory and practice of mathematical

* Corresponding author Laboratory of Environmental Chemistry and Toxicology
E-mail address: alla.toropova@marionegri.it (A.P. Toropova).

https://doi.org/10.1016/j.chemosphere.2022.137224
Received 21 September 2022; Received in revised form 3 November 2022; Accepted 9 November 2022
Available online 11 November 2022

chemistry and cheminformatics (Ambure et al., 2019), stimulating the development and improvement of new software for solving quite complex problems related to catalysis, synthesis, metabolism and drug discovery (González-Díaz et al., 2013; Halder and Cordeiro, 2021). In a sense, the antipode for classical methods involving the three-dimensional representation of molecules as well as their quantum mechanical features is the Monte Carlo method that allows you to build a model of various endpoints using the representation of the molecular structure through simplified molecular input-line entry system (SMILES) (Weininger, 1988). This approach made it possible to obtain convenient and statistically good models for a number of rather important and complex endpoints, such as Gibb's energy (Singh et al., 2022), eco-toxic endpoints of organic chemicals (Lotfi et al., 2022), toxicity of ionic liquids (Ahmadi et al., 2022), cytotoxicity of quantum dots (Kumar and Kumar, 2021a; Kumar et al., 2022), and anti-influenza activity of DNA aptamers (Kumar and Kumar, 2021b).

During the last decades, the interest in the toxicological analysis of amphibian organisms was not intensive enough, despite the clear ability of amphibians to be a valuable ecological indicator (EFSA et al., 2018; Di Nicola et al., 2021). A key factor's main reason for such a situation is the unavailability of the corresponding experimental data. Nonetheless, a few works dedicated to toxicity towards amphibian organisms are available in the literature (Wilson and Famini, 1991; Wang et al., 2001; Huang et al., 2003a,b; Roy and Ghosh, 2006; Wang et al., 2019; Toropov et al., 2022; Roy, 2022).

Here, numerical data from different sources has been collected. SMILES (Weininger, 1988) have been applied to the representation of the molecular structure. To take into account the different experimental conditions, so-called quasi-SMILES (Toropova et al., 2015) were involved. The quasi-SMILES are a convenient way to insert data of heterogeneous nature, which are not codified into the chemical structure, for instance, the exposure time. In our case, it gives the possibility to study a representative dataset on the toxic effects related to tadpoles.

The aim of the present study is the development of QSARs for toxicity towards tadpoles (*Rana japonica* and *Rana chensinensis*) using the system of self-consistent models (Toropova and Toropov, 2021) obtained by the Monte Carlo technique with the applying of a new criterion of the predictively potential of models (so-called the index of ideality of correlation) (Toropov and Toropova, 2017; Toropova et al., 2020). Thus, values on the two species are used within a single model, and this information too is used within the quasi-SMILES approach.

2. Method

2.1. Data

Experimental data on acute lethal toxicity expressed via the negative logarithm of molar concentrations pLC_{50} [mol/L] observed in 12, 24, 48, 72, and 96 h exposition of organic compounds to *Rana japonica* and *Rana chensinensis* tadpoles were extracted from the literature (Mekenyan et al., 1996; Huang et al., 2003a; Wang et al., 2019). After excluding duplicates, 188 quasi-SMILES remained. These quasi-SMILES were distributed into five non-identic combinations of four special sets: (i) active training set (≈25%); (ii) passive training set (≈25%); (iii) calibration set (≈25%); and (vi) validation set (≈25%). The total number of compounds is 52.

Table S1 (Supplementary materials) confirms that five of the above splits examined here are not identical.

Each set has a specific task. The quasi-SMILES of the active training set are applied to build the model. The quasi-SMILES of the passive training set applied to the inspection of the model (whether the model is satisfactory for quasi-SMILES that are not involved in the modelling process). Quasi-SMILES of the calibration set are applied to detect the moment of the start of the overtraining. The quasi-SMILES from the validation set are applied to final checking up the predictive potential of a model.

Quasi-SMILES were used for the representation of the molecular structure of organic chemicals as well as physicochemical and biochemical conditions of toxicological effect (Toropov et al., 2018; Toropov and Toropova, 2021). Table 1 contains the codes of the non-chemical information used within quasi-SMILES to build the models (Toropov and Toropova, 2021). In this way, the model can use data related to the exposure time and the species.

Table 1
The "cryptography" of quasi-SMILES building up.

Code for quasi-SMILES	Comment
[jap]	*Rana japonica* tadpoles
[che]	*Rana chensinesis* tadpoles
[12 h], [24 h], [48 h], [72 h], [96 h]	Time of exposition of organic compounds to tadpoles

2.2. Optimal descriptor

The CORAL model is a one-variable correlation between the endpoint and descriptor of correlation weights (*DCW*):

$$pLC_{50} = C_0 + C_1 \times DCW(T,N) \tag{1}$$

The *DCW* is a function of the molecular architecture expressed and experimental condition expressed via the quasi-SMILES

$$DCW(T,N) = \sum CW(S_k) \tag{2}$$

The S_k is SMILES atom, i.e. one symbol (e.g. 'C', 'c', 'N', 'O', etc.) or a group of symbols which cannot be examined separately (e.g. 'Cl', 'Br', %11, etc.). In addition, a special type of code is a component of the model. These reflect experimental conditions, namely [12 h], [24 h], [48 h], [72 h], and [96 h] represent the time affecting to tadpoles in 12 h, 24 h, 48 h, 72 h, and 96 h respectively; the [jap] and [che] are codes of quasi-SMILES for *Rana japonica* and *Rana chensinensis* tadpoles, respectively.

The $CW(S_k)$ is the correlation weight of the S_k, i.e. a coefficient which adds to the descriptor's value if the corresponding quasi-SMILES contains the S_k. The numerical data on the correlation weights are obtained from the Monte Carlo optimization, which is carried out with the so-called index of ideality of correlation (*IIC*), i.e. a special component of the target function described in the literature (Toropov and Toropova, 2017; Toropova et al., 2020).

2.3. Applicability domain

The applicability domain of the CORAL model is defined according to the distribution of quasi-SMILES attributes in the training and calibration sets as two-step:

Step 1: the definition of the statistical defect (d_k) for each quasi-SMILES attribute involved (non-blocked) to build up the model;

$$d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) + N(A_k)} \tag{3}$$

where $P(A_k)$ and $P'(A_k)$ are the probability of A_k in the active training and calibration sets, respectively;

$N(A_k)$ and $N'(A_k)$ are frequencies of A_k in the active training and calibration sets, respectively.

Step 2: the calculation for all quasi-SMILES statistical defect (D_j):

$$D_j = \sum_{k=1}^{NA} d_k \tag{4}$$

where NA is the number of non-blocked SMILES attributes in the SMILES.

A substance falls in the domain of applicability if

$$Dj < 2 * \overline{D} \tag{5}$$

where $\overline{D}$ is the average of the statistical SMILES-defect for the training set.

2.4. The system of self-consistent models

Each i-th model has i-th validation set. As demonstrated (Table S1), the validation sets are far from identical. It is important to develop a model which, in principle, can be used for an arbitrary validation set, provided that the substance(s) to be evaluated are similar to those of the training set. For this purpose, we can use multiple splitting of the substances available to build the model, and the results should be reproducible. In this case, these different models should be considered as self-consistent ones.

The measure of self-consistency is based on the average and dispersion of the correlation coefficients on different validation sets. The corresponding computational experiments are represented by scheme (6):

$$\begin{bmatrix} M_1 : S_1 \rightarrow R^2_{11} & \cdots & M_5 : S_1 \rightarrow R^2_{51} \\ \vdots & \ddots & \vdots \\ M_1 : S_5 \rightarrow R^2_{15} & \cdots & M_5 : S_5 \rightarrow R^2_{55} \end{bmatrix} \tag{6}$$

the M_i is i-th model (i.e. special numerical version of Eq. (1)); the S_j is j-th split; the R^2_{ij} is the correlation coefficient observed for j-th validation set if applied i-th model.

3. Results and discussion

The general model observed for random split 1 is the following:

$$pLC_{50} = 0.9771 + 0.7551 \times DCW(1, 15) \tag{7}$$

The general statistical quality of this model is the following:

n = 46, $R^2 = 0.7730$, $q^2 = 0.7557$, RMSE = 0.554, F = 150 (active training set)

n = 47, $R^2 = 0.7739$, $q^2 = 0.7570$, RMSE = 0.545, F = 154 (passive training set)

n = 48, $R^2 = 0.9537$, $q^2 = 0.9505$, RMSE = 0.292 (calibration set)

n = 47, $R^2 = 0.9697$, $q^2 = 0.9678$, RMSE = 0.262 (validation set)

n = 188, $R^2 = 0.8828$, $q^2 = 0.8809$, RMSE = 0.430 (all quasi-SMILES)

One can see (Table 2 and Table 3) that the models should be considered with good predictive potential. Table 2 contains the statistical characteristics of models on the validation set, including all compounds. In contrast, Table 3 contains the statistical characteristics of compounds which were not involved in the corresponding learning process (i.e. in the Monte Carlo optimization).

Table 3 contains the correlation weights *CW(S$_k$)* of codes of SMILES attributes as well as codes of quasi-SMILES described in Table 1. Positive values indicate factors associated with an increase in the toxic effect and vice versa. The larger coefficients indicate a higher role in the effect. Obviously, we can observe that the longer the exposure, the higher the effect. This model can be used to identify the effect depending on the time of exposure, which is an interesting characteristic of this kind of model. We can also observe a different species sensitivity between the two species; *Rana japonica* is more affected by toxic substances. This information can be used as an activity-activity relationship. As an example of the role of the chemical components, the presence of chlorine in the molecule is associated with an increased effect. The ability of the model to extract useful information is closely related to the presence of a sufficient number of substances with a certain feature. Thus, for instance, we can observe that the role of bromine in this model is very low, opposite to what was discussed for chlorine. This is undoubtedly affected by the low number of substances with bromine compared to those containing chlorine. Another critical issue related to the model, in general, is that the substances in the training set are, in practice, all aromatic ones; thus, the present model should be used for these kinds of substances.

Table 2
The determination coefficient (on the validation set) on different combinations of use of the i-th model for the j-th validation set in brackets, the numbers of compounds are indicated.

	Split 1	Split 2	Split 3	Split 4	Split 5	Average ± dispersion
The case when some of the compounds of j-th validation set are present in the training sample of the i-th model						
Model 1		0.9650 (48)	0.9212 (47)	0.9613 (47)	0.9614 (46)	0.9522 ± 0.018
Model 2	0.9757 (47)		0.9595 (47)	0.9822 (47)	0.9834 (46)	0.9752 ± 0.010
Model 3	0.9572 (47)	0.9718 (48)		0.9670 (47)	0.9708 (46)	0.9667 ± 0.006
Model 4	0.9619 (47)	0.9789 (48)	0.9471 (47)		0.9766 (46)	0.9661 ± 0.013
Model 5	0.9619 (47)	0.9778 (48)	0.9495 (47)	0.9688 (47)		0.9645 ± 0.010
The case when the compounds of j-th validation set that are present in the training set of the i-model are excluded from consideration						
Model 1		0.9878 (17)	0.9805 (13)	0.9881 (13)	0.9835 (14)	0.9850 ± 0.003
Model 2	0.9917 (17)		0.9610 (20)	0.9870 (15)	0.9870 (21)	0.9817 ± 0.012
Model 3	0.9550 (13)	0.9201 (20)		0.9555 (16)	0.9125 (18)	0.9358 ± 0.020
Model 4	0.9820 (13)	0.9839 (15)	0.9740 (16)		0.9805 (19)	0.9801 ± 0.004
Model 5	0.9777 (14)	0.9880 (21)	0.9565 (18)	0.9732 (19)		0.9739 ± 0.011

Thus, it can be seen that both the conditions for carrying out the corresponding experiments and the molecular structure features impact the toxic effect. The prevalence of the mentioned circumstances (features of molecular architecture and experimental conditions) in quasi-SMILES should be taken into account. For example, the presence of three cycles has a correlation weight of 0.35068, but the prevalence of this trait is low (Table 3). Nitrogen outside the nitro group has a negative correlation weight, while nitrogen included in the nitro group has a positive correlation weight; however, it is important that both have a significant prevalence in the active training set (N_A), passive training set N_P), and in the calibration set (N_C).

It is to be noted that the described models are quite good, but solely for the validation set. This is the consequence of applying the *IIC*: the Monte Carlo optimization based on the *IIC* improves the statistical characteristics of a model for the calibration set but to the detriment of the training (active and passive) set (Toropov and Toropova, 2017; Toropov and Toropova, 2017).

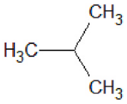
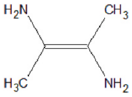
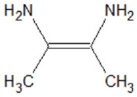
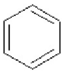
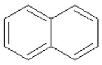
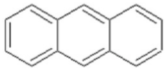
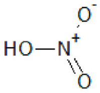
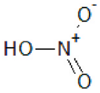
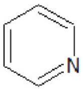
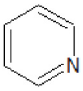
The majority of quasi-SMILES fall into the domain of applicability according to the statistical defect of SMILES (Toropova et al., 2020).

Table 4 contains the comparison of the statistical quality of models from the literature and suggested here.

One can see that the model suggested here provides quite good statistical quality for the calibration set and the validation set. The statistical quality of the CORAL model for the training set is poorer (Eq. (4)). However, the situation rather indicates Eq. (7) as a better model than the above-mentioned models from the literature, at least from the practical point of view.

Like any approach, the described scheme for using Monte Carlo has its advantages and disadvantages. The advantage of the described scheme is the relative modesty of queries regarding the initial

Table 3
The correlation weights obtained by the Monte Carlo optimization for first split.

S_k	Attribute of	Structural examples	$CW(S_k)$	N_A[a]	N_P	N_C	d_k (Eq. (3))
#	SMILES	… C≡N	−4.01232	1	2	2	0.0066
(	SMILES		0.00782	43	46	42	0.0007
/	SMILES		−0.08184	4	4	5	0.0019
\	SMILES		1.43544	4	4	5	0.0019
1	SMILES		−0.05305	46	47	43	0.0012
2	SMILES		−0.13547	5	6	11	0.0075
3	SMILES		0.35068	2	3	4	0.0066
=	SMILES	… C=N …	−0.01619	28	26	23	0.0025
C	SMILES		0.20237	14	19	16	0.0010
F	SMILES		0.0	0	1	0	0.0000
Br	SMILES		0.0	0	6	0	0.0000
Cl	SMILES		0.56788	20	22	20	0.0005
N	SMILES		−0.09010	18	16	21	0.0012
O	SMILES		0.01681	31	32	31	0.0005
[12 h]	Quasi-SMILES		0.40006	12	15	11	0.0014
[24 h]	Quasi-SMILES		0.48180	9	7	7	0.0031
[48 h]	Quasi-SMILES		0.63516	6	8	11	0.0058
[72 h]	Quasi-SMILES		0.81437	8	11	7	0.0019
[96 h]	Quasi-SMILES		0.91105	11	6	12	0.0005
[N+]	SMILES		0.68593	23	18	15	0.0049
[O-]	SMILES		0.28928	23	18	15	0.0049
[che]	Quasi-SMILES		0.07021	16	18	14	0.0019
[jap]	Quasi-SMILES		0.15024	30	29	34	0.0009
c	SMILES		0.30535	46	47	43	0.0012
n	SMILES		0.98334	2	3	4	0.0066

[a] N_A, N_P, and N_C are frequencies of quasi-SMILES in the active training set, passive training set, and calibration set, respectively.

Table 4
Comparison of the statistical quality of models for toxicity towards tadpoles.

N training	R^2 (training)	RMSE (training)	n validation	R^2 (validation)	RMSE (validation)	Hour	Tadpoles	Reference/ comment
28	0.997	0.05				24 h	*Rana japonica*	Wang et al., (2001)
9	0.93	0.22				12 h	*Rana japonica*	Wang et al., (2019)
18	0.85	0.58				96 h	*Rana chensinensis*	Wang et al., (2019)
51	0.834–0.914	0.243–0.175				12 h	*Rana japonica*	Huang et al. (2003a)
41	0.9967	0.1				48 h	*Bufo vulgaris formosus*	Yan et al., (2008)
51	0.915	0.183				12 h	*Rana japonica*	Roy and Ghosh, (2006)
44	0.722	0.330	14	0.965	0.110	12 h	*Rana japonica*	Toropov et al., (2022)
141	0.8335	0.464	47	0.9697	0.262	12, 24, 48, 72, and 96 h	*Rana japonica* and *Rana chensinensis*	In this work

information about the molecular architecture. SMILES is sufficient for the implementation of the approach. However, if experimental conditions are available (which are able to influence the endpoint), those can be presented as additional fragments of SMILES (converting SMILES in quasi-SMILES). Most of the approaches currently used are based on the choice of suitable groups of descriptors from some available source of descriptors (Wilson and Famini, 1991; Wang et al., 2001; Huang et al., 2003a,b; Roy and Ghosh, 2006; González-Díaz et al., 2013; Ambure et al., 2019; Wang et al., 2019; Halder and Cordeiro, 2021 Toropov et al., 2022; Roy, 2022). The need to evaluate many combinations of descriptors can provide certain advantages to the corresponding models, but, apparently, some complication of such models' practical applications is also inevitable. The paradoxical influence of the *IIC* (Toropov and Toropova, 2017) on the distribution of the statistical quality (increasing it for the calibration set and the external validation set to the detriment of the corresponding indicators for active training and passive training sets) is unexpected. However, this is quite an attractive situation from a practical point of view. Hence, it is not surprising models built using *IIC* find applications (Kumar and Kumar, 2021a,b; Ahmadi et al., 2022; Kumar et al., 2022; Lotfi et al., 2022; Singh et al., 2022).

Supplementary materials section contains technical details for described models.

4. Conclusions

We introduce here a new approach for modelling toxicity towards amphibians. The model takes into account heterogeneous data related to the time of exposure to a different species. This is useful both regarding the impact of the duration on the toxic effect and also from a technical point of view; this approach allows us to cope with the limited number of experimental values available because it fully exploits the values available at different times and species. The system of self-consistent models is useful to improve trust in the approach applied. The suggested method based on the index of ideality of correlation gives a quite good model for the validation set, but to the detriment of the training set. It is to be noted that the suggested approach can be adapted to QSPR/QSAR analyses for other endpoints, and the CORAL software (http://www.insilico.eu/coral) is free and available for the academic community.

Author contributions

Conceptualization, **A.P.T.**, **A.A.T.**, **A.R.**, **E.B.**; Data curation, **A.P.T.**, **A.A.T.**, **A.R.**, **E.B.**; Writing – original draft preparation, **A.P.T.**; **A.A.T.**; **A.R.**, **E.B.**; **M.R.D.N**, Writing – review & editing, **A.P.T.**; **A.A.T.**; **A.R.**, **E. B.**; **J.LC.M.D.**, **M.R.D.N.**; **J.LC.M.D**. Supervision, **A.R.**, **E.B.**; Project administration, **E.B.**; **J.L.C.M.D**. Conclusion and future perspectives, **J. L.C.M.D**. **M.R.D.N**, All authors have read and agreed to the published version of the manuscript.

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

The authors are grateful for the contribution of the project LIFE-VERMEER (contract LIFE16 ENV/IT/000167) and the EFSA project OptiTox (contract OC/EFSA/SCER/2018/01) for financial support. The views expressed in this article are the authors only and do not reflect the views of the European Food Safety Authority.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2022.137224.

References

Ahmadi, S., Lotfi, S., Kumar, P., 2022. Quantitative structure–toxicity relationship models for predication of toxicity of ionic liquids toward leukemia rat cell line IPC-81 based on index of ideality of correlation. Toxicol. Mech. Methods 32 (4), 302–312. https://doi.org/10.1080/15376516.2021.2000686.

Ambure, P., Halder, A.K., González Díaz, H., Cordeiro, M.N.D.S., 2019. QSAR-Co: an pen source software for developing robust multitasking or multitarget classification-based QSAR models. J. Chem. Inf. Model. 59 (6), 2538–2544. https://doi.org/10.1021/acs.jcim.9b00295.

Di Nicola, M.R., Cavigioli, L., Luiselli, L., Andreone, F., 2021. Anfibi & Rettili d'Italia. In: Belvedere, Latina, "historia naturae", 8, p. 576.

EFSA, P.P.R., , EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), 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. EFSA J 16 (2), 301. https://doi.org/10.2903/j.efsa.2018.5125ISSN, 5125.

González-Díaz, H., Arrasate, S., Gómez-San, A.J., Sotomayor, N., Lete, E., Besada-Porto, L., Ruso, J.M., 2013. General theory for multiple input-output perturbations in complex molecular systems. 1. linear QSPR electronegativity models in physical, organic, and medicinal chemistry. Curr. Top. Med. Chem. 13 (14), 1713–1741. https://doi.org/10.2174/1568026611313140011.

Halder, A.K., Cordeiro, M.N.D.S., 2021. QSAR-Co-X: an open source toolkit for multitarget QSAR modelling. J. Cheminf. 13 (1), 29. https://doi.org/10.1186/s13321-021-00508-0.

Huang, H., Wang, X., Ou, W., Zhao, J., Shao, Y., Wang, L., 2003a. Acute toxicity of benzene derivatives to the tadpoles (Rana japonica) and QSAR analyses. Chemosphere 53 (8), 963–970. https://doi.org/10.1016/S0045-6535(03)00715-X.

Huang, H., Wang, X., Shao, Y., Chen, D., Dai, X., Wang, L., 2003b. QSAR for prediction of joint toxicity of substituted phenols to tadpoles (Rana japonica). Bull. Environ. Contam. Toxicol. 71 (6), 1124–1130. https://doi.org/10.1007/s00128-003-8790-4.

Kumar, A., Kumar, P., 2021a. Cytotoxicity of quantum dots: use of quasi SMILES in development of reliable models with index of ideality of correlation and the consensus modelling. J. Hazard Mater. 402, 123777 https://doi.org/10.1016/j.jhazmat.2020.123777.

Kumar, P., Kumar, A., 2021b. Correlation intensity index (CII) as a benchmark of predictive potential: construction of quantitative structure activity relationship models for anti-influenza single-stranded DNA aptamers using Monte Carlo optimization. J. Mol. Struct. 1246, 131205 https://doi.org/10.1016/j.molstruc.2021.131205.

Kumar, P., Kumar, S., Kumar, A., 2022. Creation of quantitative feature toxicity relationship models for cytotoxicity of cadmium containing quantum dots towards HEK cells using quasi SMILES. Int. J. Quant. Struct.-Prop. Relatsh. 7 (1), 1–20. https://doi.org/10.4018/IJQSPR.294900.

Lotfi, S., Ahmadi, S., Kumar, P., 2022. Ecotoxicological prediction of organic chemicals toward Pseudokirchneriella subcapitata by Monte Carlo approach. RSC Adv. 12 (38), 24988–24997. https://doi.org/10.1039/d2ra03936b.

Mekenyan, O.G., Schultz, T.W., Veith, G.D., Kamenska, V., 1996. 'Dynamic' QSAR for semicarbazide-induced mortality in frog embryos. J. Appl. Toxicol. 16 (4), 355–363. https://doi.org/10.1002/(SICI)1099-1263(199607)16:4<355::AID-JAT357>3.0.CO;2-Z.

Roy, K., Ghosh, G., 2006. QSTR with extended topochemical atom (ETA) indices. VI. Acute toxicity of benzene derivatives to tadpoles (*Rana japonica*). J. Mol. Model. 12 (3), 306–316. https://doi.org/10.1007/s00894-005-0033-7.

Roy, K., 2022. Chemometrics and Cheminformatics in Aquatic Toxicology. John Wiley & Sons, p. 592. https://www.wiley.com/en- us/Chemometrics+and+Cheminformatics+in+ Aquatic+ Toxicology-p-9781119681595.

Singh, R., Kumar, P., Devi, M., Lal, S., Kumar, A., Sindhu, J., Toropova, A.P., Toropov, A.A., Singh, D., 2022. Monte Carlo based QSGFEAR: prediction of Gibb's free energy of activation at different temperatures using SMILES based descriptors. New J. Chem. 46, 19062–19072. https://doi.org/10.1039/d2nj03515d.

Toropov, A.A., Toropova, A.P., 2017. The index of ideality of correlation: a criterion of predictive potential of QSPR/QSAR models? Mutat. Res., Genet. Toxicol. Environ. Mutagen. 819, 31–37. https://doi.org/10.1016/j.mrgentox.2017.05.008.

Toropov, A.A., Toropova, A.P., Benfenati, E., Diomede, L., Salmona, M., 2018. Use of Quasi-SMILES to model biological activity of "micelle-polymer" samples. Struct. Chem. 29, 1213–1223. https://doi.org/10.1007/s11224-018-1115-3.

Toropov, A.A., Toropova, A.P., 2021. Quasi-SMILES as a basis for the development of models for the toxicity of ZnO nanoparticles. Sci. Total Environ. 772, 145532 https://doi.org/10.1016/j.scitotenv.2021.145532.

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.

Toropova, A.P., Toropov, A.A., Rallo, R., Leszczynska, D., Leszczynski, J., 2015. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. Ecotoxicol. Environ. Saf. 112, 39–45. https://doi.org/10.1016/j.ecoenv.2014.10.003.

Toropova, A.P., Toropov, A.A., Carnesecchi, E., Benfenati, E., Dorne, J.L., 2020. The using of the Index of Ideality of Correlation (IIC) to improve predictive potential of models of water solubility for pesticides. Environ. Sci. Pollut. Res. 27 (12), 13339–13347. https://doi.org/10.1007/s11356-020-07820-6.

Toropova, A.P., Toropov, A.A., 2021. The system of self-consistent of models: a new approach to build up and validation of predictive models of the octanol/water partition coefficient for gold nanoparticles. Int. J. Environ. Res. 15, 709–722. https://doi.org/10.1007/s41742-021-00346-w.

Wang, X., Dong, Y., Wang, L., Han, S., 2001. Acute toxicity of substituted phenols to Rana japonica tadpoles and mechanism-based quantitative structure-activity relationship (QSAR) study. Chemosphere 44 (3), 447–455. https://doi.org/10.1016/S0045-6535(00)00198-3.

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.

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.

Wilson, L.Y., Famini, G.R., 1991. Using theoretical descriptors in quantitative structure-activity relationships: some toxicological indices. J. Med. Chem. 34 (5), 1668–1674. https://doi.org/10.1021/jm00109a021.

Yan, D., Jiang, X., Xu, S., Wang, L., Bian, Y., Yu, G., 2008. Quantitative structure–toxicity relationship study of lethal concentration to tadpole (Bufo vulgaris formosus) for organophosphorous pesticides. Chemosphere 71, 1809–1815. https://doi.org/10.1016/j.chemosphere.2008.02.033.