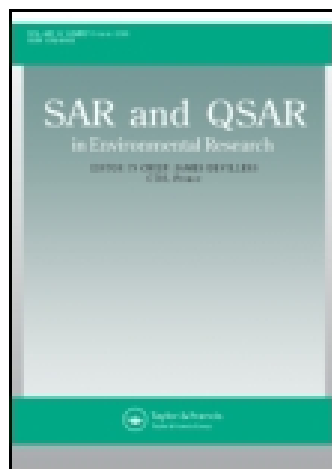


This article was downloaded by: [McMaster University]

On: 02 April 2015, At: 10:40

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



SAR and QSAR in Environmental Research

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/gsar20>

A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*)

M. Cassotti^a, D. Ballabio^a, R. Todeschini^a & V. Consonni^a

^a Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milano, Italy

Published online: 17 Mar 2015.



CrossMark

[Click for updates](#)

To cite this article: M. Cassotti, D. Ballabio, R. Todeschini & V. Consonni (2015) A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*), SAR and QSAR in Environmental Research, 26:3, 217-243, DOI: [10.1080/1062936X.2015.1018938](https://doi.org/10.1080/1062936X.2015.1018938)

To link to this article: <http://dx.doi.org/10.1080/1062936X.2015.1018938>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &

A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*)

M. Cassotti*, D. Ballabio, R. Todeschini and V. Consonni

Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milano, Italy

(Received 3 November 2014; in final form 27 January 2015)

REACH regulation demands information about acute toxicity of chemicals towards fish and supports the use of QSAR models, provided compliance with OECD principles. Existing models present some drawbacks that may limit their regulatory application. In this study, a dataset of 908 chemicals was used to develop a QSAR model to predict the LC₅₀ 96 hours for the fathead minnow. Genetic algorithms combined with *k* nearest neighbour method were applied on the training set (726 chemicals) and resulted in a model based on six molecular descriptors. An automated assessment of the applicability domain (AD) was carried out by comparing the average distance of each molecule from the nearest neighbours with a fixed threshold. The model had good and balanced performance in internal and external validation (182 test molecules), at the expense of a percentage of molecules outside the AD. Principal Component Analysis showed apparent correlations between model descriptors and toxicity.

Keywords: QSAR; fathead minnow; aquatic toxicity; REACH; similarity; *k*NN

1. Introduction

The entrance into force of REACH regulation [1] in June 2007 boosted the interest in the field of *in silico* methodologies. In fact, REACH, by introducing the concept ‘no data, no market’, obliged manufacturers and importers to prove that their products are safe for both human health and the environment. The avoidance of unnecessary testing (especially animal testing) being an explicitly declared goal, REACH provided registrants with a number of tools to pursue this objective, including the promotion of alternative test methods, such as *in vitro* and *in silico* methodologies.

Among the latter methods, quantitative structure–activity relationships (QSAR) analysis emerged as one of the suggested approaches, because of its low cost and time for application. QSAR analysis comprises a variety of mathematical and statistical methods that aim at finding functional relationships between the structure of chemical compounds, described by means of experimental or theoretical variables called molecular descriptors [2], and their measured properties and activities.

As part of the assessment of toxicity towards aquatic organisms, REACH requires the evaluation of the short-term toxic effects on fish for substances imported or manufactured in quantities greater than 10 tonnes per year (REACH Annex VIII). The benefits deriving from the availability of suitable QSAR models, both from an economical and animal welfare

*Corresponding author. Email: m.cassotti@campus.unimib.it

perspective, are evident. The idiomatic expression ‘suitable QSAR model’ indicates that the model should be scientifically valid, and the scientific validity for regulatory applications within REACH is outlined in the five principles defined by the Organization for Economic Co-operation and Development (OECD) [3]. In summary, the endpoint and the algorithm should be clearly defined, the model should be accompanied by an estimation of its domain of applicability, the goodness-of-fit and predictivity of the model should be evaluated by means of appropriate strategies and, eventually, a mechanistic interpretation of model descriptors should be given, if possible.

A number of QSAR models have been developed to predict acute toxicity towards fish, and two trends can be identified: some researchers have aimed at classifying chemicals for their mode of action (MoA) [4–6], whereas others have tackled the problem of estimating a quantitative parameter, usually the LC_{50} [7–10]. Considering quantitative models for the fathead minnow (*Pimephales promelas*), some studies focused on small homogeneous sets of chemicals belonging to the same chemical class or supposed to act via the same MoA [11–16]. The use of MoA-based QSARs for toxicity screening depends on the ability to associate query chemicals with the correct MoA, which is not an easy task. Consequently, many investigations also aimed to quantitatively model large heterogeneous datasets altogether. Mainly global strategies were employed to this end. A summary of the characteristics of quantitative models for large heterogeneous datasets is given in Table 1.

Regarding linear methods, multiple linear regression (MLR) was used in many investigations [17–30], whereas partial least squares (PLS) and the multi-linear spline regressions were seldom used [17,19,22]. On the other hand, more complex non-linear methods, such as different types of neural networks (NN) and support vector regression (SVR), were often used to model such heterogeneous datasets [20,22,23,26,31–35]. Some investigations divided chemicals into more homogeneous clusters (not necessarily corresponding to chemical classes or known MoAs) in a preliminary step and calibrated local regression models for each cluster [20,21,36–38]. The k nearest neighbours (k NN) method was also used to derive models that can be considered local because just a small neighbourhood of similar chemicals is used to estimate the toxicity of the query compound [21,39]. Read-across based on k NN was also used to assess the excess toxicity from a baseline estimated from the log P [39]. The statistics of these models, in general, were lower compared with those of models developed for specific chemical classes or modes of action.

This study focuses on the development of a new QSAR model to predict the acute toxicity of diverse chemicals, defined as LC_{50} 96 hours, towards the fathead minnow (*Pimephales promelas*). The model was developed keeping in mind the five OECD principles in order to make it applicable for regulatory purposes within REACH. To this end, attention was paid to the curation of the experimental data, which led to the definition of an extended dataset consisting of 908 organic molecules. The model, based on six molecular descriptors, used a similarity-based algorithm (k NN) to predict the toxicity. The applicability domain (AD) was automatically evaluated for each prediction and an additional analysis of the performance was carried out for individual functional groups. The predictive power was estimated by means of thorough and appropriate internal and external validation procedures. Moreover, the chemical information encoded by model descriptors was explained, and we attempted to put it in relation with aquatic toxicity. Eventually, an example of the application of the model was given.

Table 1. Characteristics of literature models for LC₅₀ towards the fathead minnow developed from large heterogeneous datasets. In case of multiple models, the range of the statistics is reported between square brackets. The dash symbol is used for lacking information.

Heterogeneous datasets									
Reference	No. models ^a	Method ^b	n train ^c	n test ^d	p ^e	r ²	Q ² _{cv}	Q ² _{ext}	
[17]	2	MLR	[560–568]	–	1	[0.61–0.65]	[0.61–0.65] ^f	–	
	20	MLR	568	–	2	[0.64–0.67]	[0.64–0.66] ^f	–	
	20	MLR	568	–	3	[0.67–0.68]	[0.67–0.68] ^f	–	
	20	MLR	568	–	4	[0.69–0.70]	[0.68–0.69] ^f	–	
	2	PLS	562	–	13	[0.72–0.73]	[0.71–0.72]	–	
[18]	1	MLR	408	57	4	0.80	0.80 ^g	0.72	
[19]	6	MLR	344	115	[6–10]	[0.76–0.79]	[0.75–0.76] ^f	[0.75–0.79]	
	3	MLR-Spline	344	115	[5–7]	[0.76–0.78]	[0.75–0.77] ^f	[0.78–0.78]	
	1	MLR	445	110	5	0.71	–	0.55	
	1	ANN	334 + 111 ^h	110	5	0.80	0.62 ⁱ	0.62	
	2	RP-MLR	445	110	[5–5] ^j	[0.75–0.76]	–	[0.60–0.63]	
[36]	8	Consensus	445	110	–	[0.78–0.80]	–	[0.63,0.67]	
[31]	1	NN + NN	454	114	156	–	–	0.76	
[39]	1	SVR	457	114	8	0.83	–	0.80	
	2	LR + kNN	692	–	–	[0.73–0.73]	[0.72–0.72] ^f	–	
	2	LR + kNN (0.8)	692 (419) ^k	–	–	[0.78–0.78]	[0.76–0.78] ^f	–	
	2	LR + kNN (0.9)	692 (230) ^k	–	–	[0.87–0.87]	[0.87–0.87] ^f	–	
	11	Fuzzy NN	392	170	9	[0.20–0.70]	–	[0.00–0.50]	
[32]	1	CPNN	275	274	150	0.97	–	0.59	
[33]	1	HC + MLR	659	164	≤ n _k /5 ^l	–	–	0.71	
[21]	1	SC + MLR	659	164	–	–	–	0.63	
	2	MLR	659	164	–	–	–	[0.69–0.70]	
	1	kNN	659	164	–	–	–	0.67	
	1	Consensus	659	164	–	–	–	0.73	
	2	MLR + NN	484	85	[10–23]	[0.82–0.75] ^m	–	[0.64–0.66] ^m	
[34]	3	MLR	607 ⁿ	–	147	[0.83–0.87]	[0.46–0.54] ⁱ	–	
[22]	3	PLS	607 ⁿ	–	147	[0.81–0.83]	[0.59–0.67] ⁱ	–	
	3	NN	607 ⁿ	–	147	[0.89–0.92]	[0.62–0.70] ⁱ	–	
	3	PMM	607 ⁿ	–	147	[0.79–0.82]	[0.48–0.60] ⁱ	–	
	1	Tree + MLR	560	–	[3–5] ^j	[0.83–0.99]	[0.81–0.98] ^{l,o}	–	

(Continued)

Table 1. (Continued).

Heterogeneous datasets								
Reference	No. models ^a	Method ^b	n train ^c	n test ^d	p ^e	r ²	Q ² _{cv}	Q ² _{ext}
[23]	1	MLR	287	88 ^p	8	0.83 ^m	0.90 ^{i,m}	0.75 ^m
[24]	2	NN	287	88 ^p	[8–8]	[0.81–0.71] ^m	[0.88–0.77] ^{i,m}	[0.84–0.74] ^m
	1	MLR	484	121	2	0.71	0.70	0.61
[38]	2	Consensus	484	121	–	–	[0.71–0.71]	[0.59–0.60]
	1	MLR + L-MLR	675	–	[2–8] ^j	0.88	0.95 ^q	–
[25]	10	MLR	557	–	[4–17]	[0.62–0.73]	[0.50–0.70] ^r	–
[26]	1	Consensus	557	201 + 144 ^s	–	0.71	–	0.60 + 0.58 ^t
	1	MLR	373	188	6	0.73	0.72 ^f	0.66
[27]	1	BPNN	373	188	6	0.78	–	0.73
	1	MLR	532	–	2	0.61	–	–
[35]	2	PNN	800	86	76	[0.89–0.99]	–	[0.78–0.52]
[28]	2	MLR	249	200	[5–6]	[0.79–0.81]	[0.78–0.80] ^g	[0.71–0.72]
[29]	3	MLR	[246–271] + [144–164] ^h	[148–158]	[1–1]	[0.67–0.68]	[0.79–0.85] ⁱ	[0.77–0.79]
[30]	1	MLR ^u						

^aNumber of developed models; ^bMLR = Multiple linear regression; PLS = Partial least squares; RP-MLR = Recursive partitioning coupled with MLR for each class; NN + NN = Clustering by means of self-organized neural network coupled with local regression by means of feedforward neural networks; SVR = Support vector regression; LR + kNN = Linear regression coupled with k nearest neighbours; LR + kNN (0.8) = Linear regression coupled with k nearest neighbours with similarity threshold = 0.8; LR + kNN (0.9) = Linear regression coupled with k nearest neighbours with similarity threshold = 0.9; Fuzzy NN = fuzzy neural network; CPNN = CounterPropagation neural network; HC + MLR = Hierarchical clustering coupled with multiple linear regression; SC + MLR = Single clustering coupled with multiple linear regression; kNN = k nearest neighbours; MLR + NN = Multiple linear regression for baseline line coupled with neural networks to model the residuals; PMM = Powell's minimization method; Tree + MLR = decision tree to partition chemicals into nine clusters coupled with MLR for each cluster; MLR + L-MLR = Multiple linear regression for baseline coupled with local MLR for clusters of molecules sharing a common toxicophore; BPNN = BackPropagation neural network; PNN = Probabilistic neural networks; ^cnumber of compounds in training set; ^dnumber of compounds in the test set; ^enumber of model descriptors; ^fleave-one-out cross-validation; ^gbootstrap with 5000 iterations; ^htraining set + validation set; ⁱQ² on the validation set; ^jnumber of descriptors in each MLR; ^knumber of compounds inside the applicability domain; ^lmaximum allowed number of descriptors in each cluster model (n_k = number of compounds in the cluster); ^mRoot mean square residuals; ⁿsplit in training-validation sets (2:1) three times; ^orange of statistics for the nine cluster models. Overall statistics not reported; ^pdivided into validation and test sets; ^qleave-10%-out cross-validation, repeated three times; ^r3-fold cross-validation; ^snumber of molecules in each of two test sets; ^tresults on each of the two test sets; ^ure-implementation in VEGA of the MLR model of [21].

2. Materials and methods

2.1 Experimental data

Experimental acute toxicities of chemicals were retrieved from three databases, namely OASIS, ECOTOX [40] and EAT5 [41]. The OASIS database was downloaded from the OECD QSAR Toolbox [42]. The databases were imported into KNIME [43] and processed by means of *ad hoc* designed workflows in order to extract the concentrations causing death in 50% of test fathead minnows (*Pimephales promelas*) over a test duration of 96 hours (LC₅₀ 96 hours). Experimental data were merged together regardless of test conditions (water pH, temperature, etc.) and test designs (flow-through, static, static renewal). In the EAT5 database, LC₅₀ data were reported as EC₅₀ (effective concentration) with lethality as observed effect. Records in the ECOTOX database indicating ranges or thresholds of experimental values were removed.

2.2 Data curation and filtering

In order to guarantee data consistency, data were checked and ambiguous molecular structures and anomalous experimental values were disregarded. Data curation and filtering were carried out in KNIME.

2.2.1 Checking identity of records

For almost every record (4626 records in total) both the CAS registry number (CAS-RN) and chemical name were available. In order to check that the CAS-RN and chemical name referred to the same structure, queries were set up to the ChemSpider database [44] and the Chemical Identifier Resolver (CIR) of the CADD Group at NCI/NIH [45]. CAS-RNs and chemical names were used independently as input for the queries. The retrieved structures were compared and if they all matched, the identity was considered correct.

Out of 4626 records (corresponding to 1139 unique CAS-RNs, plus 12 compounds lacking a CAS-RN), more than 50% presented mismatches (2422 records, corresponding to 518 different CAS-RNs and the 12 compounds lacking a CAS-RN). The records showing mismatches were exported and checked manually using PubChem [46], Sigma-Aldrich [47] and again ChemSpider as additional sources. Some records were deleted during this screening for different reasons, such as: (a) non-existent CAS-RN; (b) missing specification of which structural isomer(s) had been used; (c) unavailability of the molecular structure as it was a commercially named chemical; (d) impossibility to resolve a CAS-RN–chemical name mismatch, for example because the original publication was not found or not accessible; and (e) the record pertained to a mixture of several chemical species. At the end of this phase, 2192 records corresponding to 441 different CAS-RNs were retained and merged with the 2204 records (692 different CAS-RNs) with matching structures, giving a set of 4396 records (1060 different CAS-RNs, with 73 of the 441 CAS-RNs from the mismatching set present among the 692 CAS-RNs from the matching set).

A further inspection was carried out in order to check that each CAS-RN was associated with only one structure and vice versa. This investigation lead to the identification of 10 structures associated with two different CAS-RNs. These mismatches, probably due to obsolete CAS-RNs, were resolved by retaining only the CAS-RN indicated on the Sigma-Aldrich database.

2.2.2 Filtering and dissociation

Records with units coded as ‘%’, ‘% v/v’ and ‘AI ng/L’ were removed (13 records). The LC_{50} values of the remaining 1047 molecules were converted to molarity and transformed to logarithmic units ($-\log_{10}(\text{mol/L})$).

Several molecules had multiple experimental values, which could correspond to: (a) different measurements; (b) the same measurement published in a paper that had been included in more source databases; and (c) the same measurement published in different papers. Since the median of the LC_{50} values would have been used for modelling, duplicates of the same measurement (cases b and c) had to be removed because they would have affected the calculation of the median. Duplicates of the same experiment from different databases (case b) were removed. Duplicates of the same experiment published in more papers (case c) often lacked references. Therefore, it was decided to consider all the records with exactly the same LC_{50} value as duplicates of the same experiment and only one record was retained. This decision was taken considering that experimental measurements are implicitly affected by error, thus the probability that two measurements would give the same LC_{50} value is, in principle, extremely low.

Since the goal was to develop a model for acute toxicity limited to discrete organic molecules, only molecules with at least two carbon atoms and comprising only certain elements were retained (H, Li, B, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, and I). Symbols that specify the stereochemical configuration were removed from the *Simplified Molecular Input Line Entry System* (SMILES). Salts and mixtures were submitted to a dissociation algorithm in the OASIS Database Manager [48] that first checked whether the species could be dissociated and then screened the potential dissociation products for non-toxic species. If more than one species was considered the source of toxicity, the record was removed. By doing so, it was possible to convert 50 mixtures and salts to a single organic component, assumed as the only source of the measured toxicity. Ions such as Na^+ , Mg^{2+} , Cl^- were therefore not used for modelling. The dissociation products were neutralized, unless they were quaternary ammonium ions for which the charged form was retained. An outcome encountered for three salts was that both the organic ion (acetate, benzoate and 2-hydroxybenzoate) and the inorganic counter-ion (K^+ or Na^+) were considered not toxic by the algorithm and, consequently, removed. For these three cases, the organic component was re-introduced and considered for modelling. In 15 cases, the dissociation product coincided with another molecule in the dataset. Toxicity values of these two species were very close for most instances, thus justifying the validity of the dissociation procedure and allowing us to pool the data. In four cases, mixtures of the type A+A+B were present and the dissociation algorithm returned only one molecule of A as assumed source of toxicity. The LC_{50} (molarity) values were accordingly doubled to correct for the approximation.

At the end of this filtering stage, 929 molecules were retained. Final validation of the structures was made by comparing the SMILES in the dataset with those in the OpenTox database after processing also the latter ones with the dissociation converter. OpenTox database lacked a structure for 58 compounds in the dataset. Large agreement in the structures of the remaining 871 molecules was observed. Only nine mismatches were detected and solved by looking for the correct structure in the Sigma-Aldrich database. Only one case consisted of completely different compounds, whereas the other differences were mainly due to tautomers, valence and charge.

2.2.3 Curation of experimental values

For several compounds, multiple experimental values were available, showing differences of up to three logarithmic units. In order to reduce dependence on outlying toxicity data, the median, which is a more robust measure of central tendency than the mean, was calculated together with the corresponding standard deviation on the logarithmically transformed molarities ($-\text{Log}_{10}(\text{mol/L})$). The pooled standard deviation over the entire dataset was calculated ($\sigma = 0.229 \text{ Log}_{10}(\text{mol/L})$) and used to derive an alert for inconsistent data ($2\sigma = 0.458 \text{ Log}_{10}(\text{mol/L})$). Molecules with a standard deviation larger than 2σ were filtered out and each experimental value was searched in the original scientific publication in order to detect errors in the compilation of the databases. If the scientific publication was not available or not found, the corresponding experimental value was deleted. During this phase, 21 chemicals with large standard deviations were removed because none of the original publication was accessible or found.

The final dataset included 908 organic molecules and is freely available [49].

2.3 Molecular descriptors

The SMILES of the 908 chemicals in the dataset were used to calculate molecular descriptors by means of DRAGON 6 software [50]. Only zero-, one- and two-dimensional descriptors were calculated from the SMILES for a total of 3582 descriptors. Descriptors from the Drug-like block were not calculated because they were supposed not to be relevant for modelling aquatic toxicity. Constant, near constant and descriptors with at least one missing value were removed (1361 descriptors). Eventually, a filter on pairwise correlation was applied: if two descriptors had a coefficient of correlation greater than 0.95, only the one with the lowest average correlation with all the remaining descriptors was retained. A final pool consisting of 1218 molecular descriptors was retained and used for the subsequent modelling phase. The distribution of the 1218 retained descriptors in the 18 logical blocks of DRAGON was as follows: constitutional indices (32), ring descriptors (25), topological indices (29), walk and path counts (15), connectivity indices (14), information indices (25), 2D matrix-based descriptors (66), 2D autocorrelations (160), Burden eigenvalues (47), P_VSA-like descriptors (33), ETA indices (13), edge adjacency indices (85), functional group counts (90), atom-centred fragments (66), atom-type E-state indices (35), CATS 2D (98), 2D atom pairs (378), molecular properties (7).

2.4 Modelling methods

Since the compounds in the dataset belong to a variety of chemical classes, it is expected that they also possess different MoAs. Literature models calibrated on the largest datasets were based on either non-linear methods, or similarity-based methods, or partitioned chemicals into more homogeneous clusters for which linear models were calibrated (Table 1). This is likely to be due to the differences between MoAs. Therefore, linear modelling methods were expected not to be optimal. Among the several potentially appropriate non-linear methods, it was decided to use the k NN method because: (a) it is simple; (b) only the local neighbourhood is used to provide a prediction (presumably chemicals acting via the same MoA); and (c) it allows a chemical analysis for each test molecule and its nearest neighbours.

The distance of a molecule from all the molecules in the training set is computed and the k training molecules with the lowest distances are selected. The experimental response values

of the k closest training molecules, i.e. the nearest neighbours, are used to calculate the prediction. Obviously, in fitting, the distance of a molecule with itself is neglected. The Jaccard–Tanimoto distance was used to find the nearest neighbours. The Jaccard–Tanimoto distance between two molecules r and t , d_{rt} , was derived from the corresponding coefficient as [51]:

$$d_{rt} = \left(1 - \frac{\sum_{j=1}^p x_{rj} \cdot x_{tj}}{\sum_{j=1}^p x_{rj}^2 + \sum_{j=1}^p x_{tj}^2 - \sum_{j=1}^p (x_{rj} \cdot x_{tj})} \right)^{1/2} \quad (1)$$

$$= \left(1 - \frac{d_{rt,euclidean}^2}{\sum_{j=1}^p x_{rj}^2 + \sum_{j=1}^p x_{tj}^2 - \sum_{j=1}^p (x_{rj} \cdot x_{tj})} \right)^{1/2} \quad 0 \leq d_{rt} \leq 1$$

where j runs over the p variables. Then, the prediction \hat{y}_r for molecule r was taken as the weighted mean over the k nearest neighbours, where the weights were calculated as a function of the distance, according to Equation (2):

$$\hat{y}_r = \sum_{t=1}^k y_t \cdot w_t = \sum_{t=1}^k y_t \cdot \frac{(1 - d_{rt})}{\sum_{t=1}^k (1 - d_{rt})} \quad (2)$$

where y_t and w_t are the experimental response and the weight of the t th neighbour, respectively, and the sum runs over the k neighbours. Molecular descriptors were scaled in the range [0,1] prior to computing the distances.

The k NN method was combined with genetic algorithms (GAs) in order to select the relevant molecular descriptors.

2.5 Applicability domain assessment

A two-step procedure was implemented in order to have an in-depth assessment of the AD of the model. A preliminary bounding box approach is carried out prior to finding the nearest neighbours. Test compounds with descriptors values outside the range of the training set are therefore considered outside the AD. For all the compounds that have descriptors values within the range of the training set (and are therefore considered inside the AD by the bounding box approach), the further evaluation of the AD is based on the distance with the nearest neighbours as described below.

A number of k NN similarity-based approaches have been defined in the scientific literature [52–54] to assess the AD of QSAR models. These methods are based on the calculation of a similarity measure of the molecule to be predicted with respect to molecules in the training set. The similarity can be calculated from the distances of the molecule to be predicted from its nearest neighbours. The obtained distance (or similarity) measure is then compared with a user-defined threshold. If the distance is within the threshold, the query molecule is considered to have enough similar neighbours to assure a reliable prediction: the molecule falls inside the AD of the model. With respect to other basic AD approaches (such as bounding box), k NN AD approaches better describe the molecules distribution because they can locally describe the covariance structure of the data. Therefore, k NN AD approaches better define the QSAR model space but, obviously, there may be training molecules considered outside the AD, while a bounding box method would include all training molecules.

In this study we used the same approach previously applied to predict acute toxicity towards *D. magna* [55]. In summary, for each molecule the average distance from its k nearest neighbours taken from the training set was compared with a fixed distance threshold. If the average distance was greater than the distance threshold, the molecule to be predicted

and its k nearest neighbours were regarded to as relatively dissimilar, and therefore the molecule was considered outside the AD. On the contrary, if the average distance was lower than the distance threshold, the molecule was considered to be enough similar to its neighbours (in the training set) to allow a reliable prediction and said to be inside the AD of the model. In this way, the evaluation of the AD is implicitly defined on a k NN similarity-based approach and carried out on the fly for each prediction.

Since the model is not parametric and is based on local similarities, compounds from the training set that did not have sufficiently similar neighbours were still retained in the training set because they could be useful for the prediction of future test compounds.

2.6 Model validation

In order to properly validate the model, the dataset of 908 compounds was randomly divided into a training set (726 chemicals) and a test set (182 molecules). The training set was used to carry out variable selection by means of GAs, following the strategy proposed by Leardi and González [56], and calibrate the final model. The settings used for GAs are reported in Table 2.

During GA runs, the performance of the models was assessed by means of internal five-fold cross-validation with venetian blinds splitting of the training samples. The coefficient of determination in cross-validation (Q_{cv}^2), defined according to Equation (3), was used as fitness function.

$$Q_{cv}^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

where y_i and \hat{y}_i are the experimental and predicted responses of the i th object, respectively; \bar{y} is the average response value. Genetic algorithms were used to select molecular descriptors, optimize the number of nearest neighbours (k) and the distance threshold used for the assessment of the AD. Values of k (number of nearest neighbours) from 1 to 10 were tried for each model.

In order to carry out a more thorough validation, the final model was also internally validated by means of leave-more-out strategy: 20% of training molecules, randomly selected, were left out. The procedure was reiterated 1000 times and the average coefficient of determination (Q_{cv}^2) was calculated.

Table 2. Settings used for genetic algorithms.

Option	Value	option	Value
number of chromosomes	30	twins allowed	no
average number of variables in the chromosomes of starting population	5	hybridization	yes
mutation probability	0.01	frequency of hybridization	1 every 100 runs
cross-over probability	0.50	number of cross-validation groups	5
number of independent runs	100	cross-validation type	venetian blinds
number of evaluations for each run	100	maximum number of nearest neighbours	10

Finally, the test set was used to assess the predictive power of the model calibrated on the training set. The performance on the test set was assessed by means of the Q^2_{F3} function [57], defined as:

$$Q^2_{ext} = 1 - \frac{PRESS/n_{ext}}{TSS/n_{tr}} = 1 - \frac{\left[\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2 \right] / n_{ext}}{\left[\sum_{i=1}^{n_{tr}} (y_i - \bar{y})^2 \right] / n_{tr}} \quad (4)$$

where n_{tr} and n_{ext} are the number of molecules in the training and test sets, respectively.

2.7 Software

KNIME [43] was used to extract the relevant data from the source databases and process them by means of *ad hoc* designed workflows. The OASIS Database Manager [48] was used to retain only organic compounds, apply the dissociation converter and compare the SMILES in the dataset and in the OpenTox database. DRAGON 6 [50] was used to calculate molecular descriptors and apply unsupervised variable reduction. MATLAB [58] was used to carry out variable selection and model validation by means of routines written by the authors. Marvin was used for drawing, displaying and characterizing chemical structures and substructures [59]. ChemProp [60] was used to retrieve the molecules used for the application example (paragraph 3.4).

3. Results and discussion

3.1 Model development and analysis

3.1.1 Variable selection and model calibration

Variable selection was carried out on the training set of 726 chemicals in subsequent steps in order to handle the large number of calculated descriptors, i.e. 1218, and avoid potential overfitting. First, GAs were run separately on each block of DRAGON descriptors (18 blocks in total). The number of independent runs (from which the selection frequencies were calculated) was set to 100. Then, only the descriptors with the largest frequencies of selection from each block were retained and merged together to form a pool of 208 candidate good descriptors, which was input to GA again. Results based on 100 independent runs showed that only one molecular descriptor, i.e. *MLOGP* [61,62], had a considerably larger frequency of selection than the others. Models based on all the possible combinations of the 15 most frequently selected descriptors were calculated with the constraint that *MLOGP* be always included, since it was demonstrated to be relevant for toxicity modelling. The obtained models were then judged taking into consideration both their predictive power and their complexity, which was defined by the number of included descriptors and their ease of interpretation. This procedure resulted in a *k*NN model (*k* equal to six) based on six molecular descriptors (*MLOGP*, *CIC0*, *NdssC*, *NdsCH*, *SM1_Dz(Z)*, *GATS1i*).

3.1.2 Definition of the applicability domain

After selecting the optimal set of descriptors and the number of nearest neighbours (*k* equal to six), an analysis was carried out in order to define the optimal value of the distance threshold, which is used in the second step of the assessment of the AD, as explained previously.

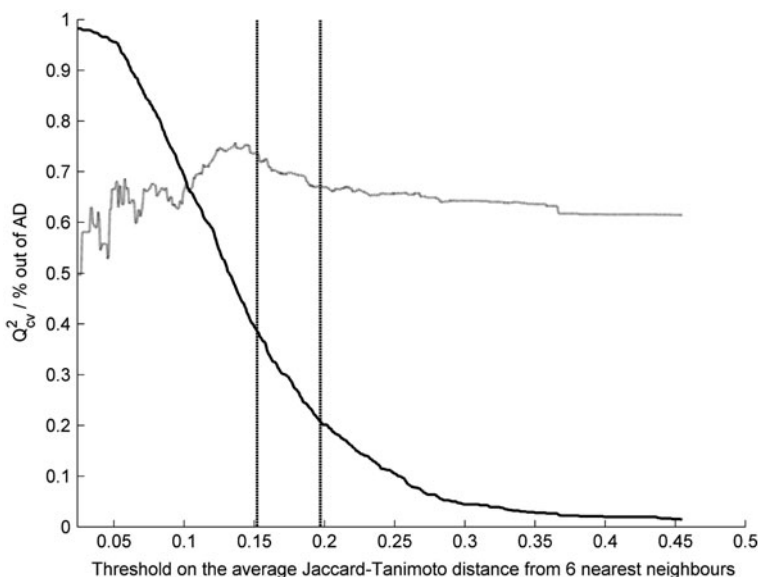


Figure 1. Q^2_{cv} (dotted line) and percentage of molecules out of AD (solid line) as a function of the threshold value on the average Jaccard–Tanimoto distance from six neighbours. The two vertical lines are the selected distance thresholds.

Since only compounds with average distance from the six nearest neighbours lower than the threshold are considered inside the AD, it is evident that a low distance threshold corresponds to a strict AD criterion because it demands all six neighbours to be very similar (low distance) to the molecule to be predicted. Figure 1 shows the Q^2_{cv} values and the percentage of molecules outside the AD of the model, as a function of the distance threshold value. From Figure 1 it can be noted that very low distance threshold values correspond to large percentages of molecules out of AD (as expected), but surprisingly the Q^2_{cv} values are not high. The performance of the model (Q^2_{cv}) rises with the distance threshold value up to a maximum and then decreases smoothly, whereas the percentage of chemicals out of AD always decreases more steeply. Two distance thresholds were chosen, corresponding to values of 0.152 and 0.197, which will be referred to as the ‘Strict’ and ‘Soft’ distance thresholds, respectively. The ‘Strict’ distance threshold (0.152) is located nearby the maximum performance of the model.

The chemicals considered inside the AD with the ‘Strict’ distance threshold are a subset of those inside the AD with the ‘Soft’ threshold. This subset is characterized by lower distance (higher similarity) between the compound to be predicted and its nearest neighbours and higher prediction accuracy (as shown in the following paragraph). Thus, with the ‘Strict’ distance threshold, the percentage of compounds out of AD is larger, but the model performs better. It should be stressed that the predictions provided by the model with the ‘Soft’ and ‘Strict’ distance thresholds are the same. The ‘Soft’ and ‘Strict’ distance thresholds can be used to fulfil different needs. The ‘Strict’ distance threshold is intended to be used when the risk of using a potentially low-accuracy prediction is high and it is therefore preferable to have no prediction at all. On the other hand, the ‘Soft’ distance threshold can be applied to situations where it is more desirable to have a toxicity estimate (even if potentially less

accurate) rather than having none, e.g. for high-throughput screening or for regulatory applications in the framework of a weight of evidence approach. In any case, the user can tune the value of the distance threshold to fulfil personal needs.

3.1.3 Model statistics and analysis of the residuals

The statistics of the model in fitting, cross-validation (both five-fold and leave-more-out) and external validation corresponding to both chosen threshold values ('Strict' and 'Soft' distance thresholds) are shown in Table 3. The model reached satisfactory performance both in internal and external validation, especially considering the size of the dataset (to our knowledge the largest analysed so far). Moreover, the performances obtained with both distance thresholds ('Soft' and 'Strict') in fitting, internal and external validation, are balanced, which should indicate absence of overfitting, a common pathology when dealing with high-dimensional data. As expected, the 'Soft' threshold has lower (yet still satisfactory) statistics compared with the 'Strict' threshold, but the percentage of molecules outside the AD is nearly half. If only the bounding box approach is used to evaluate the AD, i.e. no distance threshold, the statistics would be significantly lower (r^2 , Q^2_{cv} and Q^2_{ext} equal to 0.62, 0.61 and 0.61, respectively), as highlighted in Table 3 ('No' threshold case), thus indicating that a polyhedral-like representation of the training set space is not appropriate. On the contrary, the similarity-based evaluation of the AD seems effective in identifying unreliable predictions when used in the framework of a k NN model.

Figures 2(a) and 2(b) show the calculated and predicted *versus* experimental toxicity values for the training set (fitting) and the test set, respectively. Compounds inside the AD with the 'Strict' distance threshold are represented by star symbols. The vast majority of these compounds are accurately predicted. The few associated with poor toxicity estimates are affected by both underestimation and overestimation. If the 'Soft' distance threshold is used, also molecules represented by the addition symbol are regarded inside the AD. It is apparent that some of these molecules are not very well predicted (again both overestimated and underestimated). This behaviour seems more evident in the training set (Figure 2(a)). Nevertheless, many of the molecules associated with the worst predictions are regarded as outside the AD with both the 'Soft' and 'Strict' distance threshold. The model seems not to have a noticeable bias.

Figures 2(c) and 2(d) report the standardized residuals *versus* the average Jaccard–Tanimoto distance from the six nearest neighbours for the training and test sets, respectively. It can be seen that the residuals tend to increase when increasing the average distance, i.e. molecules with larger average distances from their neighbours (less similar) are associated with less accurate predictions. This fact is the experimental justification of the ideas behind the introduction of the similarity-based AD approach: the more structurally similar (low distance) a molecule to its nearest neighbours, the more similar their responses and, therefore, the more reliable the prediction. Nevertheless, two opposite behaviours that do not follow these assumptions can be detected. On one side, there are molecules associated with large average distances (out of AD) that are instead well predicted. This could be the case of structural cliffs, i.e. molecules with relatively different structures that possess instead similar activities. On the other side, there are molecules with relatively low average distances, i.e. similar, that are poorly predicted. This could instead be the case of activity cliffs, i.e. molecules with similar structures but different activities. In this regard, there are two chemicals in the training set that are inside the AD of the 'Strict' distance threshold (star symbols) but whose predictions are very poor (absolute standardized residuals larger than three). This situation occurs for an

Table 3. Statistics of the model in fitting, cross-validation and external validation. Training set: 726 molecules; test set: 182 molecules.

Threshold	k	thr	Fitting			5-fold CV			Leave-more-out			Test set		
			r ²	RMSEC	% out ^a	Q ² _{cv}	RMSEC	% out ^a	Q ² _{cv}	RMSEC	% out ^a	Q ² _{ext}	RMSEP	% out ^a
'No' ^b	6	—	0.62	0.879	0	0.61	0.878	1	0.61	0.890	1	0.61	0.888	1
'Soft'	6	0.197	0.69	0.752	17	0.67	0.755	21	0.72	0.756	20	0.73	0.745	15
'Strict'	6	0.152	0.73	0.657	33	0.74	0.641	38	0.79	0.654	39	0.77	0.682	27

^aPercentage of molecules outside the applicability domain of the model; ^bonly bounding box assessment of the applicability domain.

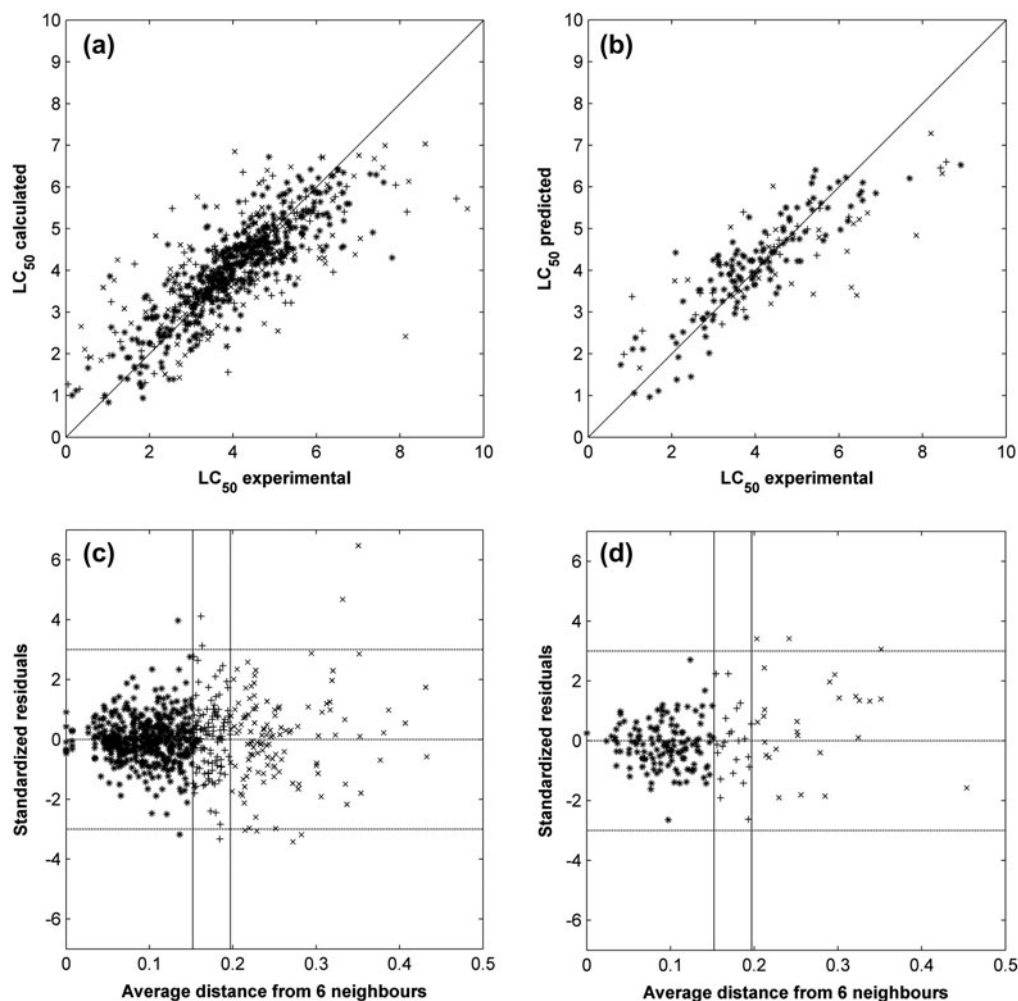


Figure 2. Results with the 'Strict' and 'Soft' distance thresholds: a) calculated vs. experimental LC_{50} for training set; b) predicted vs. experimental LC_{50} values for test set; standardized residuals vs. average Jaccard-Tanimoto distance from six neighbours for training set (c) and test set (d). Multiplication symbols: molecules out of AD with both distance thresholds; star symbols: molecules in AD with both distance thresholds; addition symbols: molecules in AD of 'Soft' and out of 'Strict' distance thresholds. Vertical lines in Figures 2(c) and 2(d) correspond to the distance threshold values. (LC_{50} values are reported as $-\log_{10}(\text{mol/L})$).

additional three compounds in the training set with the 'Soft' distance threshold, while none of the test set molecules inside the AD with either the 'Soft' or 'Strict' distance threshold has absolute standardized residuals greater than three. These five training chemicals with absolute standardized residuals higher than three are provided in Table 4 together with additional details. Two of these compounds are pyrethroid insecticides and their toxicity is largely underestimated. *N*-vinylcarbazole is predicted less toxic than it is, as well, and had also been detected as an outlier elsewhere [17]. The comparison of the experimental toxicity of these three chemicals with the baseline toxicity calculated by means of the equation reported in

Schüürmann et al. [39] indicates that they exert excess toxicity. The MoA of *N*-vinylcarbazole and Flucythrinate was determined to be electrophile/pro-electrophile and CNS seizure agent, respectively [4]. The LC_{50} ($-\log_{10}(\text{mol/L})$) of the remaining two compounds (3,3-dimethylglutaric acid and bis(2-ethylhexyl) phthalate) is instead overestimated. Also, their calculated baseline toxicities are greater than the experimental values. From these results, it can be hypothesized that the model has a tendency to underestimate the toxicity of pyrethroids. An investigation, indeed, highlighted that all pyrethroids in the dataset are underestimated. As aforementioned, the AD assessment approach is based on the idea that similar molecules possess similar toxicities: the predicted toxicity of a test molecule with enough similar neighbours is assumed reliable. Unfortunately, presumed reliable predictions are not always accurate, as for the five analysed molecules, because the reliability is evaluated only on the chemical structures, but the accuracy depends on the model performance, which is not uniform in the chemical structural domain.

3.1.4 Performance on individual functional groups

An additional analysis of the performance of the model with the ‘Strict’ threshold value was carried on individual functional groups. For the list of moieties included in DRAGON software, the root mean square error (RMSE) was calculated from the results in five-fold cross-validation only on the molecules that feature a specific functional group. To this end, the number of occurrences of a functional group within one chemical obtained from DRAGON was transformed into a binary value indicating presence or absence. It should be highlighted that combinations of functional groups were not considered. The results are displayed in Figure 3, which is a bubble plot where the size of each bubble is proportional to the number of molecules that possess each functional group; the *y*-axis is the RMSE between experimental and predicted responses and the *x*-axis ranks the functional groups from the least to the most represented. Figure 3(a) shows that the error of the model on infrequent moieties (left-side) varies significantly from very low, e.g. for (thio/dithio) sulfonates (*nSO3*), aliphatic oximes (*nRCNO*) and aliphatic compounds with secondary or tertiary sp^2 carbon atoms (*nR=Ct* and *nR=Cs*), to very high, e.g. pyrrolidines (*nPyrrolidines*), pyrroles (*nPyrroles*), (thio/dithio) sulfonic acids (*nSO2OH*) and dihalogenated sp^3 carbon atoms (*nCR2X2*). However, drawing conclusions from the results on such infrequent functional

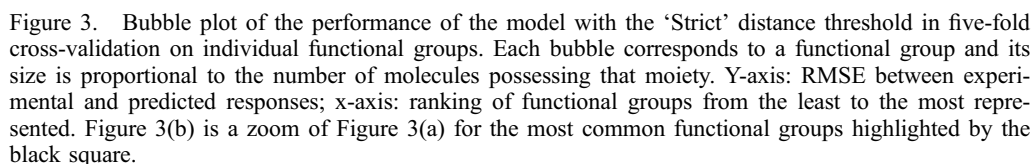
Table 4. Details of the five training chemicals with absolute standardized residuals greater than 3 inside the AD of the ‘Strict’ or ‘Soft’ distance threshold. All LC_{50} values are expressed as $-\log_{10}(\text{mol/L})$.

Name	CAS-RN	$pLC_{50}^{exp^a}$	$pLC_{50}^{pred^b}$	Class ^c	$pLC_{50}^{narc^d}$	AD ^e
3,3-Dimethylglutaric acid ^f	4839-46-7	1.055	3.859	dicarboxylic acid	2.079 ^k	‘Strict’‘Soft’
<i>N</i> -Vinylcarbazole ^g	1484-13-5	7.809	4.305	carbazole	4.228 ^k	‘Strict’‘Soft’
Bis(2-ethylhexyl) phthalate ^h	117-81-7	2.548	5.485	phthalate	7.870	‘Soft’
Fenpropathrin ⁱ	39515-41-8	8.169	5.406	pyrethroid	6.255	‘Soft’
Flucythrinate ^j	70124-77-5	9.354	5.722	pyrethroid	6.680	‘Soft’

^aexperimental LC_{50} value; ^bpredicted LC_{50} value; ^cchemical class; ^dbaseline toxicity according to [39];

^eindicates the distance thresholds for which the chemical is inside the applicability domain; ^fOC(=O)CC(C(=O)O)(C)C; ^gC=C1c2ccccc2c1cccc2; ^hCCCCC(COC(=O)c1ccccc1C(=O)OCC(CCCC)CC)CC;

ⁱN#CC(c1cccc(c1)Oc1ccccc1)OC(=O)C1C(C1(C)C)(C)C; ^jN#CC(c1cccc(c1)Oc1ccccc1)OC(=O)C(c1ccc(cc1)OC(F)F)C(C)C; ^k*MLOGP* used because no experimental log P was found.



groups would be misleading because the small number of molecules having these moieties does not make a reliable statistical sample. The RMSE values converge along the x -axis to values close to the average RMSE over the entire dataset (0.641) because more represented functional groups are considered. Figure 3(b) allows to make some considerations regarding the most common functional groups. The majority of moieties associated with large RMSE comprise compounds with hydroxyl groups or sp^3 carbon atoms: in particular, secondary and tertiary alcohols (*'nOHs'* and *'nOht'*), tertiary and quaternary sp^3 carbon atoms (*'nCt'* and *'nCq'*). *'nROH'* accounts in general for hydroxyl groups including secondary and tertiary alcohols. The performance on primary alcohols (*'nOHp'*) is instead much better: in fact, a trend of increasing RMSE values for the sequence primary < secondary < tertiary alcohols (*'nOHp'* – *'nOHs'* – *'nOht'*) is evident. The same applies also to the sequence primary < secondary < tertiary < quaternary sp^3 carbon atoms (*'nCp'* – *'nCq'* – *'nCt'* – *'nCq'*). In some cases, the performance of the model on the aromatic form is worse than on the corresponding aliphatic, e.g. primary amines (*'nArNH2'* versus *'nRNH2'*) and ethers (*'nArOR'* versus *'nROR'*). Other functional groups show instead the opposite trend, e.g. alcohols (*'nArOH'* versus *'nROH'*) and esters (*'nArCOOR'* versus *'nRCOOR'*). The error associated with molecules with conjugated π systems (*'nCconj'*) is greater than that associated with aromatic compounds (*'nCar'*). The same applies to thioethers (*'nRSR'*) that have a greater RMSE than ethers (*'nROR'*). The performance on aromatic tertiary amines (*'nArNR2'*) is slightly worse than on primary amines (*'nArNH2'*). The same consideration applies also to hydrogen bond donors and acceptors (*'nHDon'* and *'nHAcc'*, respectively). No difference seems to exist between substituted and unsubstituted benzene C atoms (*'nCbb'* and *'nCbbh'*, respectively), probably because they are often present in the same compound. It should be highlighted that the RMSEs of 'broad functional groups' that include in their definition more specific ones

(e.g. *nHAcc*, *nHDon*) are an average of the RMSEs of the functional groups converging into them. The fact that the frequent moieties associated with the largest RMSE values correspond to higher degrees of substitution (secondary and tertiary alcohols, tertiary and quaternary sp^3 carbon atoms) could indicate that the model tends to perform less well on chemicals featuring branches.

The aim of this analysis is to provide users with additional information regarding the performance of the model that can be considered to assess the reliability of each prediction. Therefore, the comments outlined above should be taken as indications. Detailed information regarding the performance of the model on individual functional groups in fitting, five-fold cross-validation and external validation with both the 'Soft' and 'Strict' distance threshold is provided in the supplementary material, available via the Supplementary Content tab on the article's online page at <http://dx.doi.org/10.1080/1062936X.2015.1018938>.

3.2 Interpretation of model descriptors

The proposed *k*NN model is based on six molecular descriptors (*MLOGP*, *CIC0*, *NdssC*, *NdsCH*, *SMI_Dz(Z)*, *GATS1i*) selected by means of GAs. Here we give a description of model descriptors, together with a coarse-grained interpretation of their relationship with fish toxicity.

MLOGP is the octanol-water partitioning coefficient ($\log P$) calculated by means of the Moriguchi model, which consists in a regression equation based on 13 structural parameters [61,62]. The $\log P$ is a widely accepted estimate of the lipophilicity of organic compounds, which is considered the driving force of narcosis.

CIC0 belongs to the set of indices of neighbourhood symmetry [63]. These indices derive from a partitioning of the vertices of the hydrogen-filled molecular graph into equivalency classes. According to this scheme, two vertices (i.e. atoms) are equivalent if they represent the same chemical element and their neighbourhood of order k , i.e. the bonded atoms up to topological distance equal to k , is identical. In particular, *CIC0* is the complementary information index of order zero, i.e. only graph vertices (i.e. atoms) are considered: it is calculated as deviation of the information content of order zero (*IC0*) from its maximum value. The value of this index decreases with increasing number of different chemical elements present in a molecule: thus, it can be said to encode information regarding heteroatoms, where only the number of different elements is accounted for and not the number of occurrences of each element. Information content indices, including *CIC* indices, have already been proven useful in biological correlations in general [64], and more specifically for modelling acute toxicity to *Pimephales promelas* of alcohols [11] and esters [12].

NdssC and *NdsCH* belong to the atom-type E-state counts, a simplification of the Kier–Hall atom-type E-state indices [65,66] in that they only count the number of occurrences of given atom-types [67]. In particular, *NdssC* and *NdsCH* count the number of unsaturated sp^2 carbon atoms of the type $=\text{C}<$ and $=\text{CH}-$, respectively. Hence, these two descriptors account for a variety of functional groups with double bonds, e.g. (thio)ketones and aldehydes, imines, carboxylic and carbamic acids, amides, esters and carbon–carbon double bonds. A common characteristic is an electrophilic carbon atom that can react with nucleophiles, giving substitution or addition reactions. At the borderline we find carbon–carbon double bonds because they can give addition reactions in which the double bond first acts as nucleophile, generating an intermediate electrophilic carbocation, which is then attacked by another nucleophile (e.g. hydrohalogenation). The hypothesis that these descriptors encode information about the electrophilic characteristics of chemicals is corroborated by the presence of several nucleophiles in living organisms and the fact that electrophiles are among the most common

toxicants. A further indication of the appropriateness of *NdsCH* can be found in In et al. [20], where the corresponding index calculated as sum of the E-states, *SdsCH*, had been used in a decision tree to classify reactive chemicals from narcotics. *NdsCH* and *NdssC* are, indeed, related to the electrophiles/pro-electrophiles MoA.

SM1_Dz(Z) belongs to a set of descriptors calculated from 2D matrices derived from the molecular graph (2D matrix-based descriptors) [2]. In particular, *SM1_Dz(Z)* is the spectral moment of order 1 calculated from the Barysz matrix weighted by the atomic number [68]. In other words, this descriptor is the sum of the eigenvalues of the Barysz matrix, whose elements take into account information on both the bond order and the atomic number. Also this descriptor seems to account for heteroatoms. The largest observed correlation is, in fact, with the number of heteroatoms ($\rho = 0.86$): the molecules with the lowest *SM1_Dz(Z)* values are entirely constituted by carbon atoms (both aromatic and not) while the largest values are taken on by highly fluorinated and chlorinated compounds and, more in general, compounds with several heteroatoms.

GATS1i is a 2D Geary autocorrelation descriptor [2]. Geary coefficients vary from zero to infinite and assume low values for positive autocorrelations and vice versa. In particular, *GATS1i* considers the ionization potential of atom pairs at topological distance equal to one, i.e. bonded atoms. *GATS1i* tends to have low values for molecules with pairs of bonded atoms with comparable ionization potentials, such as CC and CBr. Consequently, *GATS1i* tends to have low values for (a) molecules with several carbon-carbon bonds, as highlighted by the relatively large coefficient of correlation with the percentage of carbon atoms ($\rho = -0.79$), and (b) molecules with bromine and iodine. In addition, the distribution of *GATS1i* acquires lower values for aromatic compounds.

Since the model is based on a *k*NN approach, there are no coefficients to quantify the contribution of each descriptor in the calculation of toxicity. The analysis of how descriptors relate to toxicity was carried out by means of principal component analysis [69]. The score and loading plots of the training set are shown in Figure 4. An evident trend in the toxicity values emerges in Figure 4(a): toxicity increases mainly from right to left along PC1; a minor increase is also observed from top to bottom along PC2. Two descriptors have high loadings on PC1, namely *MLOGP* and *GATS1i*, and two on PC2, i.e. *CIC0* and *SM1_Dz(Z)* (Figure 4(b)). The score plot in Figure 4(a) shows that molecules with larger *MLOGP*, i.e.

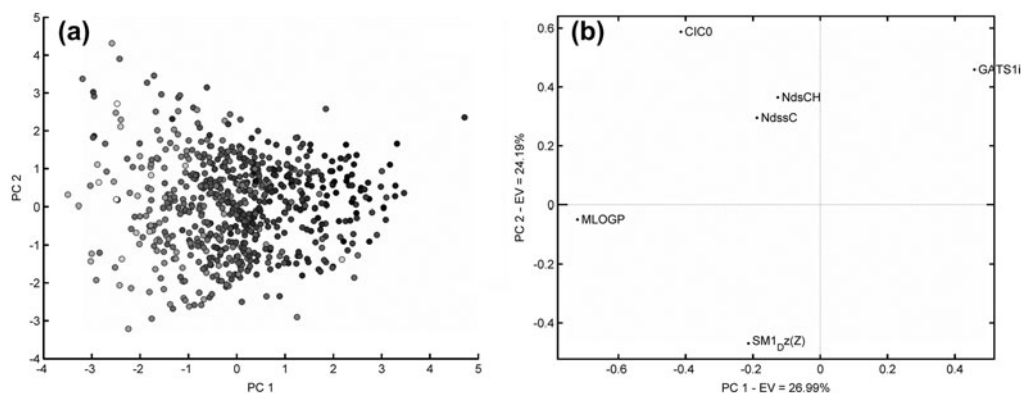


Figure 4. Principal component analysis of the training set. a) score plot with molecules coloured based on the toxicity values. White: high toxicity; black: low toxicity. b) loading plot of model descriptors.

more lipophilic, tend to have greater toxicity. It is widely known that baseline toxicity is strictly connected with the partitioning of xenobiotics between water and organism, which in turn relies on lipophilicity. The octanol-water partition coefficient has been widely used as an estimate of lipophilicity and is present in most equations to calculate narcosis-level toxicity. The second contribution is provided by *GATS1i*. In this case, toxicity increases with decreasing values of the descriptor. It was previously noticed that low values of *GATS1i* are taken on by molecules with high carbon content, among which especially aromatic compounds are found. This information seems also to be related to lipophilicity. Thus, *GATS1i* seems to account for similar information as log P, at least in regard of the main observed trend along PC1, which is related to lipophilicity. Toxicity also slightly increases moving downwards along PC2. As aforementioned, *CIC0* and *SM1_Dz(Z)* account for the presence of heteroatoms, with which *CIC0* and *SM1_Dz(Z)* have an inverse and direct correlation, respectively. From Figure 4(a) it can be seen that higher toxicity is possessed by molecules with lower *CIC0* and larger *SM1_Dz(Z)* values, i.e. more heteroatoms. This trend might be due to specific reactions, which can vary with the MoA. The remaining two descriptors, *NdssC* and *NdsCH*, have low loadings on PC1 and relatively low ones on PC2. As aforementioned, these descriptors account for the presence of specific electrophilic functional groups: the role of *NdssC* and *NdsCH* is supposed to be limited to finding similar neighbours for compounds having such moieties, i.e. electrophiles/pro-electrophiles.

3.3 Comparison with existing models

Published models to predict the LC₅₀ towards the fathead minnow developed from large heterogeneous datasets have been based on a variety of modelling methods and descriptors (Table 1). The performance of global models in cross-validation ranged from values of Q^2_{cv} equal to 0.46 [22] up to 0.85 [29] with MLR, while SVR gave the best results in external validation (Q^2_{ext} equal to 0.80 [31]). The largest statistics in cross-validation (Q^2_{cv} equal to 0.87) were obtained by a similarity-based assessment of the AD in a model that combined global and local techniques [39]. ‘Local’ MLR models calibrated on individual clusters of chemicals gave higher statistics in internal validation [37,38]. However, these models lacked an external validation to test the whole procedure of clustering and toxicity prediction. The following detailed analysis will focus on literature models calibrated on the largest datasets for the sake of comparison with our model.

Niculescu et al. achieved high statistics in fitting and external validation (Q^2_{ext} equal to 0.78) on the second largest dataset by means of NN [35]. The high statistics obtained on all test chemicals indicate high accuracy, but the model lacks an approach to estimate the AD in which this accuracy is granted.

The third largest dataset was modelled in the T.E.S.T. and VEGA software [21,30]. Good predictivity was provided by the T.E.S.T. *consensus* model (Q^2_{ext} equal to 0.73) and the MLR VEGA model with a narrow AD (Q^2_{ext} equal to 0.69).

The model of Schüürmann et al. combined the concept of baseline toxicity with read-across to evaluate the toxicity enhancement [39]. High statistics in fitting and leave-one-out internal validation (Q^2_{cv} equal to 0.87) were obtained by a strict AD criterion, but no external validation was carried out.

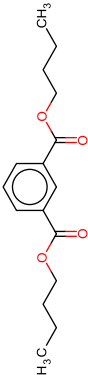
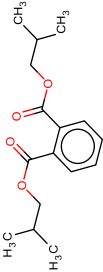
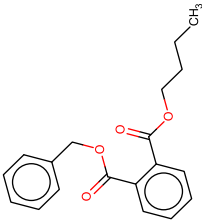
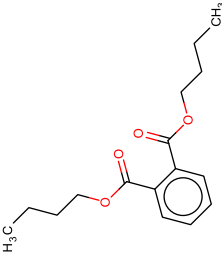
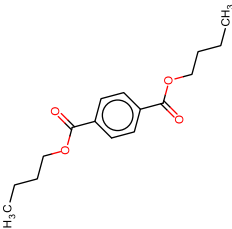
The remaining QSAR models were calibrated on smaller datasets, often using the sole MED-Duluth database. Some authors used complex regression methods, such as NN and SVR.

The model proposed in this study was calibrated on a dataset that, to our knowledge, is the largest published so far: 908 compounds. This implied a higher structural diversity and therefore, presumably, presented additional challenges for modelling. Data were collected from different sources and presented high variability for the same chemical. This aspect rendered the calibration of QSAR models more difficult compared with using data measured in the same laboratory. The performance of the model is comparable with those of models in the literature, especially regarding the predictivity on the test set. In fact, the highest accuracy in prediction (Q^2_{ext} equal to 0.80 [31]) is similar to that achieved by our model with the ‘Strict’ criterion on a larger test set (Q^2_{ext} equal to 0.77). The model is based on a reduced number of descriptors (six), in contrast to some literature models [22,33,35,36]. In addition, the six descriptors are derived from the simple 2D structure. Some published models were instead based on 3D and quantum-chemical descriptors [17,18,23,25–28,31–33,36,37] that required geometry optimization, which can be time-consuming and might also limit the future application of the model due to inconsistencies with the generation of 3D structures. The model is also built with a simple k NN algorithm based on local similarities. This aspect has two beneficial effects: from the modelling viewpoint, it can handle non-linearity and is supposed to overcome the issue related to the different modes of action because only the local neighbourhood participates in the prediction; from the regulatory application viewpoint, simple algorithms are more transparent and therefore provide increased confidence in their use. In contrast, some literature models were based on more complex algorithms and strategies, such as SVR [31], NN [20,22,23,26,32–36] or introduction of a preliminary classification/clustering step [20,21,36–38]. In addition, the model implements a systematic AD assessment, which is lacking in several literature models. Eventually, in compliance with OECD principle four, the model was thoroughly validated by means of appropriate techniques (five-fold and leave-more-out cross-validation and external validation). This should assure that the reported statistics are reasonably valid for real applications of the model. On the other hand, the predictive power of some published models was assessed by means of less strict procedures, e.g. lack of an external validation [17,22,27,37–39] or internal validation by means of leave-one-out strategy [17,19,26,37,39], a technique that was reported to give optimistic results on large datasets [70,71]. Considering all these points, the model presented in this study may be considered satisfactory.

3.4 Application example

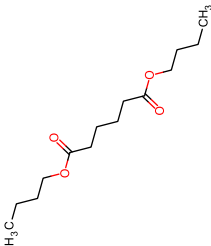
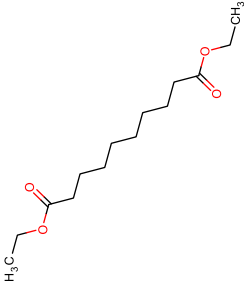
An example of the application of the model to external molecules is provided in this section. The results presented here do not constitute a validation of the model, but only serve the purpose of showing the information provided in the output and the further analyses that can be undertaken by the user in order to evaluate the reliability of each prediction. The dataset of the model published in Russom et al. [4] for predicting the MoA was retrieved from its implementation in the ChemProp software [60]. Nine molecules were unambiguously identified as being external to the dataset of the model and were therefore submitted for prediction. Five molecules, namely tetrabutyltin, chloroform, chloromethyl styrene, dichloromethane and iodoform, were outside the AD of the model with the ‘Strict’ distance threshold because the average distance from the six nearest neighbours was greater than the distance threshold (0.152). The predictions of the LC_{50} ($-\text{Log}_{10}(\text{mol/L})$), for the four molecules inside the AD, namely tetraethyltin, di-*n*-butylisophthalate, 4,9-dithiadodecane and *p*-chlorophenyl-*o*-nitrophenyl ether, were equal to 4.41, 5.28, 4.98 and 5.44, respectively, *versus* experimental values equal to 7.33, 5.49, 4.84, and 5.11, respectively. It is evident that the predictions for

Table 5. Example of the output of the model for di-*n*-isobutylphthalate. All LC_{50} values are expressed as $-\text{Log}_{10}(\text{mol/L})$.

TEST			Average distance ^e	Distance threshold	Applicability domain
Name	CAS-RN	Yexp ^a	5.28	0.057	in AD
di- <i>n</i> -butylisophthalate	3126-90-7	5.49			
Identified functional groups	RMSEP ^d				
Functional Group			RMSEP ^d		
Primary C (sp ³)	0.713		0.785		
aromatic C	0.660		0.659		
substituted benzene C	0.659		1.081		
aromatic ester	0.439		0.714		
				0.152	
					
NEIGHBOUR 2					
Name	diisobutyl phthalate				
CAS-RN	84-69-5				
Yexp ^a	5.49				
Std. dev ^e	—				
No. data ^f	1				
Ycalc ^g	5.03				
Distance ^h	0.000				
Ref. ⁱ	[40]				
					
NEIGHBOUR 4					
Name	benzyl butyl phthalate				
CAS-RN	85-68-7				
Yexp ^a	5.22				
Std. dev ^e	0.13				
No. data ^f	2				
Ycalc ^g	5.66				
Distance ^h	0.086				
Ref. ⁱ	[40]				
					
NEIGHBOUR 3					
Name	di- <i>n</i> -butyl phthalate				
CAS-RN	84-74-2				
Yexp ^a	5.37				
Std. dev ^e	0.23				
No. data ^f	8				
Ycalc ^g	5.03				
Distance ^h	0.000				
Ref. ⁱ	[40][41]				
					
NEIGHBOUR 1					
Name	di- <i>n</i> -butyl terephthalate				
CAS-RN	1962-75-0				
Yexp ^a	5.67				
Std. dev ^e	0.010				
No. data ^f	2				
Ycalc ^g	5.03				
Distance ^h	0.000				
Ref. ⁱ	[40][41]				
					

(Continued)

Table 5. (Continued).

NEIGHBOUR 5		NEIGHBOUR 6	
Name	dibutyl adipate	Name	diethyl decanedioate
CAS-RN	105-99-7	CAS-RN	110-40-7
Yexp ^a	4.85	Yexp ^a	4.98
Std. dev ^e	—	Std. dev ^e	0.017
No. data ^f	1	No. data ^f	3
Ycalc ^g	5.04	Ycalc ^g	5.04
Distance ^h	0.128	Distance ^h	0.128
Ref. ⁱ	[40]	Ref. ⁱ	[40]
			

^aexperimental LC₅₀; ^bpredicted LC₅₀; ^caverage distance from nearest neighbours; ^dRMSEP of the model on molecules in the test set that have a specific functional group; ^estandard deviation and number of experimental values from which the median used in the training set was calculated; ^fcalculated LC₅₀ in fitting; ^hJaccard–Tanimoto distance between test–training molecule; ⁱsource of experimental values from which the median used in the training set was calculated.

di-*n*-butylisophthalate, 4,9-dithiadodecane and *p*-chlorophenyl-*o*-nitrophenyl ether are accurate and affected by an error that is lower than the RMSEP obtained on the external validation (0.682, Table 3). The prediction of tetraethyltin is instead affected by a large error, probably due to the lack of compounds with tin atoms in the training set. Therefore, a warning about the potential low accuracy of this prediction could be derived from the composition of the training set. Chloroform, iodoform and dichloromethane are correctly regarded out of AD; in fact, only molecules with at least two carbon atoms were retained in the training set.

In order to give a practical example of the use of the model, the prediction of di-*n*-butylisophthalate is discussed and the output of the model is provided in Table 5. Table 5 gives information regarding the test molecule in terms of predicted value, average distance from the six neighbours (which was compared with the distance threshold) and the outcome of the assessment of the AD. Furthermore, the test molecule was screened against a list of functional groups. The results of the screening report the list of identified moieties and corresponding RMSEP of the model (on the test set) in order to provide insight about the performance of the model on test molecules bearing the same moieties. In addition, detailed information regarding the nearest neighbours is given. Name, CAS-RN and structure allow to specify the identity of the nearest neighbours and visually evaluate the similarity among the nearest neighbours and between each neighbour and test molecule. The median LC₅₀ value of the neighbours used in the training set ('Yexp') is given, together with information about the number of experimental values used for its calculation ('No. data') and the corresponding standard deviation ('Std. dev'). Reference to the source of the data is provided in the 'Ref' field. These fields allow two analyses to be carried out: on one side, it is possible to check whether there is a large variability in the experimental values for each individual neighbour molecule (undesirable situation) and, on the other side, allows checking whether all the neighbours have similar toxicity values (desired situation) or not. The calculated LC₅₀ for the neighbours ('Ycalc') is also provided in order to gain knowledge about the performance of the model in the analysed area of the chemical space. Accurate model estimates in the neighbourhood should enhance the confidence in the prediction for the test molecule as well. Finally, the Jaccard–Tanimoto distance between test molecule and each neighbour is given.

In the analysed case of di-*n*-butylisophthalate, the similarity with the neighbours is very high. In fact, all the neighbours are esters of dicarboxylic acids (four benzenedioates and two linear aliphatic dioates). The range of experimental LC₅₀ values for the neighbours is limited to less than one log unit [4.85–5.67]. The standard deviations of the experimental values of the neighbours are also in general low, which enhances the confidence in their accuracy. The situation depicted in this example shows high homogeneity between test molecule and neighbours, but it can be reasonably expected that structures that are more heterogeneous could be present among the neighbours in other cases.

4. Conclusions

This study addressed the problem of predicting the acute toxicity (LC₅₀ after 96 hours) of chemicals towards the fathead minnow (*Pimephales promelas*) by means of a QSAR model that can be used in the REACH regulatory framework.

Toxicity data from different sources were analysed and led to the definition of a large dataset that was modelled altogether. The *k*NN method was used to estimate the toxicity. The similarity among chemicals was evaluated from six molecular descriptors that do not require geometry optimization.

The AD was assessed by a systematic procedure that is applied to each chemical to be predicted by a two-step procedure. The comparison of the average distance from the nearest neighbours with a distance threshold seemed effective in describing the distribution of chemicals in the model space and identifying unreliable predictions. The distance threshold can be changed to tune the strictness of the AD criterion.

The model was thoroughly validated both internally and externally by means of a test set. Considering the size of the dataset, the variability of the experimental data taken from different sources, the simplicity of the algorithm and the low number of molecular descriptors, the model achieved satisfactory performance. Comparison with literature models showed the statistics to be comparable with those of models calibrated on large (yet smaller) datasets.

Evident correlations between model descriptors and toxicity were highlighted. The main trends were associated with the effect of lipophilicity and number of heteroatoms. Three descriptors were related to known modes of action. Since the dataset was modelled altogether, it was expected that descriptors related to more general trends, rather than closely to each MoA, would be selected.

Eventually, an amount of information regarding the nearest neighbours and the performance of the model on individual functional groups can be provided to the users to further assess the reliability of each prediction.

Acknowledgements

The authors wish to acknowledge Eva Bay Wedeby and Nikolai Georgiev Nikolov from the Technical University of Denmark, National Food Institute for helping with the curation of the dataset and making software packages and the OpenTox database available.

References

- [1] *Regulation (EC) No 1907/2006*. 2006, pp. 1–849.
- [2] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2nd ed., Vol. 41, Wiley-VCH, 2009.
- [3] OECD. The Organization for Economic Development and Co-operation, *Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models*, The Organization for Economic Development and Co-operation, ENV/JM/MONO(2007)2, 2007; available at [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono\(2007\)2](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2).
- [4] C.L. Russom, S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond, *Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas)*, Environ. Toxicol. Chem. 16 (1997), pp. 948–967.
- [5] L. Michielan, L. Pireddu, M. Floris, and S. Moro, *Support vector machine (SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals*, Mol. Inform. 29 (2010), pp. 51–64.
- [6] M. Nendza, M. Müller, and A. Wenzel, *Discriminating toxicant classes by mode of action: 4. Baseline and excess toxicity*, SAR QSAR Environ. Res. 25 (2014), pp. 393–405.
- [7] A. Levet, C. Bordes, Y. Clément, P. Mignon, H. Chermette, P. Marote, C. Cren-Olivé, and P. Lantéri, *Quantitative structure-activity relationship to predict acute fish toxicity of organic solvents*, Chemosphere 93 (2013), pp. 1094–1103.
- [8] W.D. Marzio and M.E. Saenz, *Quantitative structure-activity relationship for aromatic hydrocarbons on freshwater fish*, Ecotoxicol. Environ. Saf. 59 (2004), pp. 256–262.
- [9] K. Rose and L.H. Hall, *E-state modeling of fish toxicity independent of 3D structure information*, SAR QSAR Environ. Res. 14 (2003), pp. 113–129.

- [10] G. Tugcu, M.T. Saçan, M. Vracko, M. Novic, and N. Minovski, *QSTR modelling of the acute toxicity of pharmaceuticals to fish*, SAR QSAR Environ. Res. 23 (2012), pp. 297–310.
- [11] S.C. Basak and V.R. Magnuson, *Molecular topology and narcosis. A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC)*, Arzneim.-ForschungDrug Res. 33 (1983), pp. 501–503.
- [12] S.C. Basak, D.P. Gieschen, and V.R. Magnuson, *A quantitative correlation of the LC50 values of esters in Pimephales promelas using physicochemical and topological parameters*, Environ. Toxicol. Chem. 3 (1984), pp. 191–199.
- [13] L.D. Newsome, D.E. Johnson, R.L. Lipnick, S.J. Broderius, and C.L. Russom, *A QSAR study of the toxicity of amines to the fathead minnow*, Sci. Total Environ. 109 (1991), pp. 537–551.
- [14] G.D. Veith and S.J. Broderius, *Structure-toxicity relationships for industrial chemicals causing type (II) narcosis syndrome*, in *QSAR in Environmental Toxicology*, K.L.E. Kaiser, ed., Springer, The Netherlands, 1987, pp. 385–391.
- [15] G.D. Veith, D.J. Call, and L.T. Brooke, *Structure-toxicity relationships for the fathead minnow, Pimephales promelas: Narcotic industrial chemicals*, Can. J. Fish. Aquat. Sci. 40 (1983), pp. 743–748.
- [16] A.P. Bearden and T.W. Schultz, *Structure-activity relationships for Pimephales and Tetrahymena: A mechanism of action approach*, Environ. Toxicol. Chem. 16(1997) (1997), pp. 1311–1317.
- [17] T.I. Netzeva, A.O. Aptula, E. Benfenati, M.T.D. Cronin, G. Gini, I. Lessigarska, U. Maran, M. Vračko, and G. Schüürmann, *Description of the electronic structure of organic chemicals using semiempirical and ab initio methods for development of toxicological QSARs*, J. Chem. Inf. Model. 45 (2005), pp. 106–114.
- [18] M. Pavan, T.I. Netzeva, and A.P. Worth, *Validation of a QSAR model for acute toxicity*, SAR QSAR Environ. Res. 17 (2006), pp. 147–171.
- [19] K. Roy and R.N. Das, *QSTR with extended topochemical atom (ETA) indices. 15. Development of predictive models for toxicity of organic chemicals against fathead minnow using second-generation ETA indices*, SAR QSAR Environ. Res. 23 (2012), pp. 125–140.
- [20] Y.-Y. In, S.-K. Lee, P.-J. Kim, and K.-T. No, *Prediction of acute toxicity to fathead minnow by local model based QSAR and global QSAR approaches*, Bull. Korean Chem. Soc. 33 (2012), pp. 613–619.
- [21] T. Martin, P. Harten, R. Venkatapathy, and D. Young, *T.E.S.T. (Toxicity Estimation Software Tool)*. U.S. E.P.A., 2012; available at <http://www.epa.gov/nrmrl/std/qsar/qsar.html>.
- [22] M. Casalegno, E. Benfenati, and G. Sello, *An automated group contribution method in predicting aquatic toxicity: The diatomic fragment approach*, Chem. Res. Toxicol. 18 (2005), pp. 740–746.
- [23] D.V. Eldred, C.L. Weikel, P.C. Jurs, and K.L.E. Kaiser, *Prediction of fathead minnow acute toxicity of organic compounds from molecular structure*, Chem. Res. Toxicol. 12 (1999), pp. 670–678.
- [24] M. Hewitt, M.T.D. Cronin, J.C. Madden, P.H. Rowe, C. Johnson, A. Obi, and S.J. Enoch, *Consensus QSAR Models: Do the benefits outweigh the complexity?*, J. Chem. Inf. Model. 47 (2007), pp. 1460–1468.
- [25] S. Lozano, M.-P. Halm-Lemeille, A. Lepailleur, S. Rault, and R. Bureau, *Consensus QSAR related to global or MOA models: Application to acute toxicity for fish*, Mol. Inform. 29 (2010), pp. 803–813.
- [26] U. Maran, S. Sild, P. Mazzatorta, M. Casalegno, E. Benfenati, and M. Romberg, *Grid computing for the estimation of toxicity: Acute toxicity on fathead minnow (Pimephales promelas)*, in *Distributed, High-Performance and Grid Computing in Computational Biology*, W. Dubitzky, A. Schuster, P. Sloot, M. Schroeder, and M. Romberg, eds., Springer, Berlin, 2007, pp. 60–74.
- [27] M. Nendza and C.L. Russom, *QSAR modelling of the ERL-D fathead minnow acute toxicity database*, Xenobiotica 21 (1991), pp. 147–170.
- [28] E. Papa, F. Villa, and P. Gramatica, *Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow)*, J. Chem. Inf. Model. 45 (2005), pp. 1256–1266.

- [29] A.P. Toropova, A.A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, and G. Gini, *Coral: QSAR models for acute toxicity in fathead minnow (Pimephales promelas)*, J. Comput. Chem. pp. 1218–1223.
- [30] *VEGA Non-Interactive Client*. Milano, Italy: Istituto di Ricerche Farmacologiche Mario Negri; available at <http://www.vega-qsar.eu/index.php>.
- [31] Y. Wang, M. Zheng, J. Xiao, Y. Lu, F. Wang, J. Lu, X. Luo, W. Zhu, H. Jiang, and K. Chen, *Using support vector regression coupled with the genetic algorithm for predicting acute toxicity to the fathead minnow*, SAR QSAR Environ. Res. 21 (2010), pp. 559–570.
- [32] P. Mazzatorta, E. Benfenati, C.-D. Neagu, and G. Gini, *Tuning neural and fuzzy-neural networks for toxicity modeling*, J. Chem. Inf. Model. 43 (2003), pp. 513–518.
- [33] P. Mazzatorta, M. Vracko, A. Jezierska, and E. Benfenati, *Modeling toxicity by using supervised kohonen neural networks*, J. Chem. Inf. Model. 43 (2003), pp. 485–492.
- [34] J. Devillers, *A new strategy for using supervised artificial neural networks in QSAR*, SAR QSAR Environ. Res. 16 (2005), pp. 433–442.
- [35] S.P. Niculescu, A. Atkinson, G. Hammond, and M. Lewis, *Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow*, SAR QSAR Environ. Res. 15 (2004), pp. 293–309.
- [36] G. Gini, M.V. Craciun, C. Konig, and E. Benfenati, *Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity*, J. Chem. Inf. Model. 44 (2004), pp. 1897–1902.
- [37] A. Colombo, E. Benfenati, M. Karelson, and U. Maran, *The proposal of architecture for chemical splitting to optimize QSAR models for aquatic toxicity*, Chemosphere 72 (2008), pp. 772–780.
- [38] G. Klopman, R. Saiakhov, and H.S. Rosenkranz, *Multiple computer-automated structure evaluation study of aquatic toxicity II. Fathead minnow*, Environ. Toxicol. Chem. 19 (2000), pp. 441–447.
- [39] G. Schüürmann, R.-U. Ebert, and R. Kühne, *Quantitative read-across for predicting the acute fish toxicity of organic compounds*, Environ. Sci. Technol. 45 (2011), pp. 4616–4622.
- [40] US EPA. U.S. Environmental Protection Agency, *ECOTOX Database, Release 4.0*; available at <http://cfpub.epa.gov/ecotox/>.
- [41] ECETOC. European Centre for Ecotoxicology and Toxicology Of Chemicals, *TR 091 - ECETOC Aquatic Toxicity (EAT) database*. 2003; available at <http://www.ecetoc.org/technical-reports>.
- [42] *The OECD QSAR Toolbox for Grouping Chemicals into Categories*. Organisation for Economic Co-operation and Development, 2010; available at <http://www.qsartoolbox.org/>.
- [43] M.R. Berthold, N. Cebon, F. Dill, T.R. Gabriel, T. Kötter, T. Meinel, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, *KNIME: The Konstanz Information Miner*, in *Studies in Classification, Data Analysis, and Knowledge Organization*, 2007, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker, eds., Springer Berlin Heidelberg, pp. 319–326.
- [44] Royal Society of Chemistry, *ChemSpider*; available at <http://www.chemspider.com>.
- [45] NCI/CADD Group, *Chemical Identifier Resolver*; available at <http://cactus.nci.nih.gov>.
- [46] E.E. Bolton, Y. Wang, P.A. Thiessen, and S.H. Bryant, *PubChem: Integrated platform of small molecules and biological activities*, Annu. Rep. Comput. Chem. 4 (2008), pp. 217–241.
- [47] Sigma-Aldrich Co.; available at <http://www.sigmaaldrich.com>.
- [48] N. Nikolov, V. Grancharov, G. Stoyanova, T. Pavlov, and O. Mekenyan, *Representation of chemical information in OASIS centralized 3D database for existing chemicals*, J. Chem. Inf. Model. 46 (2006), pp. 2537–2551.
- [49] MICHEM. Milano Chemometrics and QSAR Research Group, *Acute toxicity to fish dataset*; available at <http://michem.disat.unimib.it/chm/download/toxicityfish.htm>.
- [50] DRAGON 6 (*Software for Molecular Descriptor Calculation*). Talete srl, 2012.
- [51] R. Todeschini, D. Ballabio, and V. Consonni, *Distances and other dissimilarity measures in chemometrics*, in *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Accepted for publication.
- [52] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, and R. Todeschini, *Comparison of different approaches to define the applicability domain of QSAR models*, Molecules 17 (2012), pp. 4791–4810.

- [53] F. Sahigara, D. Ballabio, R. Todeschini, and V. Consonni, *Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions*, J. Cheminformatics 5 (2013), p. 27.
- [54] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, and S.K. Kearsley, *Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR*, J. Chem. Inf. Model. 44 (2004), pp. 1912–1928.
- [55] M. Cassotti, D. Ballabio, V. Consonni, A. Mauri, I.V. Tetko, and R. Todeschini, *Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN Method*, ATLA — Altern. Lab. Anim. 42 (2014), pp. 31–41.
- [56] R. Leardi and A.L. González, *Genetic algorithms applied to feature selection in PLS regression: How and when to use them*, Chemom. Intell. Lab. Syst. 41 (1998), pp. 195–207.
- [57] V. Consonni, D. Ballabio, and R. Todeschini, *Comments on the definition of the Q₂ parameter for QSAR validation*, J. Chem. Inf. Model. 49 (2009), pp. 1669–1678.
- [58] *MATLAB*. Natick, MA, USA: MathWorks Inc., 2012.
- [59] *Marvin*. ChemAxon Ltd., 2012.
- [60] Chemical Properties Estimation Software System, *ChemProp*, UFZ Department of Ecological, Chemistry, Leipzig, 2013.
- [61] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, and Y. Matsushita, *Simple method of calculating octanol/water partition coefficient*, Chem. Pharm. Bull. 40 (1992), pp. 127–130.
- [62] I. Moriguchi, S. Hirono, I. Nakagome, and H. Hirano, *Comparison of reliability of log P values for drugs calculated by several methods*, Chem. Pharm. Bull. 42 (1994), pp. 976–978.
- [63] V.R. Magnuson, D.K. Harriss, and S.C. Basak, *Topological indices based on neighborhood symmetry: Chemical and biological applications*, in *Studies in Physical and Theoretical Chemistry*, R.B. King, ed., Elsevier, Amsterdam, The Netherlands, 1983, pp. 178–191.
- [64] S.C. Basak, D.K. Harriss, and V.R. Magnuson, *Comparative study of lipophilicity versus topological molecular descriptors in biological correlations*, J. Pharm. Sci. 73 (1984), pp. 429–437.
- [65] L.H. Hall and L.B. Kier, *Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information*, J. Chem. Inf. Model. 35 (1995), pp. 1039–1045.
- [66] L.H. Hall, L.B. Kier, and B.B. Brown, *Molecular similarity based on novel atom-type electrotopological state indices*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 1074–1080.
- [67] D. Butina, *Performance of Kier-Hall E-state descriptors in quantitative structure activity relationship (QSAR) studies of multifunctional molecules*, Molecules 9 (2004), pp. 1004–1009.
- [68] M. Barysz, G. Jashari, R.S. Lall, A.K. Srivastava, and N. Trinajstić, *On the distance matrix of molecules containing heteroatoms*, in *Chemical Applications of Topology and Graph Theory*, R.B. King, ed., Elsevier, Amsterdam, The Netherlands, 1983, pp. 222–230.
- [69] I. Jolliffe, *Principal component analysis*, in *Encyclopedia of Statistics in Behavioral Science*, B. S. Everitt and D.C. Howell, eds., John Wiley & Sons, Ltd, 2005.
- [70] A. Golbraikh and A. Tropsha, *Beware of q²!*, J. Mol. Graph. Model. 20 (2002), pp. 269–276.
- [71] K.H. Esbensen and P. Geladi, *Principles of proper validation: Use and abuse of re-sampling for validation*, J. Chemom. 24 (2010), pp. 168–187.