Prediction of Acute Aquatic Toxicity Toward *Daphnia magna* by using the GA-*k*NN Method

Matteo Cassotti,[1] Davide Ballabio,[1] Viviana Consonni,[1] Andrea Mauri,[1] Igor V. Tetko[2,3,4] and Roberto Todeschini[1]

[1]University of Milano-Bicocca, Department of Earth and Environmental Sciences, Milano, Italy; [2]Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Munich, Germany; [3]Chemistry Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia; [4]eADMET GmbH, Garching, Germany

Summary — In this study, a QSAR model was developed from a data set consisting of 546 organic molecules, to predict acute aquatic toxicity toward *Daphnia magna*. A modified *k*-Nearest Neighbour (*k*NN) strategy was used as the regression method, which provided prediction only for those molecules with an average distance from the *k* nearest neighbours lower than a selected threshold. The final model showed good performance (R^2 and Q^2_{cv} equal to 0.78, Q^2_{ext} equal to 0.72). It comprised eight molecular descriptors that encoded information about lipophilicity, the formation of H-bonds, polar surface area, polarisability, nucleophilicity and electrophilicity.

Key words: *aquatic toxicity,* Daphnia magna*, genetic algorithms,* k*NN, QSAR*.

Address for correspondence: *Davide Ballabio, University of Milano-Bicocca, Department of Earth and Environmental Sciences, Piazza della Scienza 1, Milano 20126, Italy.*
E-mail: *davide.ballabio@unimib.it*

Introduction

Many chemicals partition in water and can exert adverse effects on aquatic systems, damaging aquatic species and food webs, and threatening the survival of other members of these ecosystems, such as birds and mammals (1). The adverse effects of toxicants can be induced by means of both non-specific and specific mechanisms of action. Non-specific interactions, e.g. narcosis and general reactivity, derive from high concentrations of the toxicants within the cell or cellular membrane, and thus are strongly related to the ability of chemicals to enter the organism.

Some chemicals are able to directly interact with biological targets within the aquatic organism, causing higher toxicity (compared to the baseline set by narcosis). These interactions, or reactions, usually take place between the toxicant (or its metabolites) and critical cellular macromolecules.
The assessment of the aquatic toxicity of chemicals is a primary aspect to be addressed. Toxicity tests are typically divided into acute and chronic tests (2), according to the duration of the exposure. Information about the acute aquatic toxicity of chemicals is required for all substances subject to the European Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation (3). In particular, Annex VII of REACH suggests that *Daphnia* is used as the preferred organism for short-term aquatic toxicity testing. REACH promotes the use of alternative test methods, such as *in vitro* and computer-based methods, including Quantitative Structure–Activity Relationships (QSAR; 3), which are mathematical models that relate the structure of chemical compounds to their activities/properties by using molecular descriptors (4). The Organisation for Economic Co-operation and Development (OECD; 5) set five principles that should be fulfilled by a QSAR model, in order for it to be applicable for regulatory purposes.

Several QSAR models which address acute toxicity toward *D. magna* were calibrated both on heterogeneous and homogeneous data sets, the latter comprising only one specific class of chemical. A list of published QSAR models is reported in Table 1. In general, QSAR models developed on homogeneous (6–14) data sets had higher performances than models calibrated on heterogeneous data. When dealing with QSAR models calibrated on large heterogeneous data sets (8, 15–19), model statistics are lower than those of models calibrated on homogeneous data sets. This is probably due to non-linearity introduced by different mechanisms of action.

To the best of our knowledge, three published QSAR models demonstrated good performance on large heterogeneous data sets. Kaiser (17) developed four QSAR models by using Probabilistic Neural Networks (PNN) coupled with linear corrections with 57 molecular descriptors, calibrated on 700

Table 1: Published QSAR models for acute toxicity toward *D. magna*

Reference	Chemical class	No. of models	*n* training	*n* test	R^2	p	Q^2_{cv}	Q^2_{ext}
Homogeneous data sets								
Vighi (6)	Organophosphorus	1	22	—	0.89	6	—	—
Vighi (7)	Organotin	14	< 15	—	[0.44–0.99]	[1–3]	—	—
Todeschini (8)	Amines	1	8	—	1.00	4	1.00	—
Todeschini (8)	Chlorobenzenes	1	6	—	1.00	3	1.00	—
Todeschini (8)	Organotin	1	15	—	0.99	6	0.97	—
Todeschini (8)	Organophosphorus	1	20	—	0.92	5	0.85	—
Deneer (9)	Nitroaromatics	3	[15–22]	—	[0.60–0.75]	[1–2]	—	—
Hossain (10)	Ionic liquids	1	64	—	0.97	25	—	—
Zvinavashe (11)	Organothiophosphate	3	10	5	[0.80–0.82]	[1–2]	[0.62–0.73]	[0.61–0.71]
Cassani (12)	Triazoles and benzotriazoles	2	97	NR	[0.73–0.77]	[5–5]	[0.70–0.74]	[0.68–0.83]
Cassani (13)	Triazoles and benzotriazoles	7	90	—	[0.59–0.82]	[5–245]	[0.70–0.75]	—
Tetko (14)	Polybrominated diphenyl ethers	1	46	—	0.70	150	0.70	—
Heterogeneous data sets								
Todeschini (8)	—	5	49	—	[0.68–0.82]	[3–7]	[0.64–0.74]	—
Faucon (15)	—	1	61	35	0.54	2	0.49	0.57
Katritzki (16)	—	1	86	44	0.70	5	0.64	0.74
Katritzki (16)	—	2	87	43	[0.72–0.78]	5	[0.67–0.75]	[0.66–0.54]
Katritzki (16)	—	1	130	—	0.71	5	0.68	—
Kaiser (17)	—	4	700	76	[0.87–0.88]	57	—	[0.76–0.76]
Kar (18)	—	1	222	75	0.69	7	0.68	0.74
Kühne (19)	—	1	1365	—	0.85	NR	0.84	—

p = number of molecular descriptors in the model. NR = not reported; Q^2_{cv} = coefficient of determination in cross-validation; Q^2_{ext} = coefficient of determination in external validation; R^2 = coefficient of determination in fitting.
Bibliographic reference, chemical class (where relevant), number of developed models, number of molecules in training and external test sets are reported. In the case of multiple models, the range of the statistics is reported in square brackets.

training compounds and externally validated with 76 molecules (Q^2_{ext} equal to 0.76). Kar (18) collected experimental data on 297 chemicals, and the best QSAR model was obtained by using Partial Least Squares (PLS) regression with seven molecular descriptors (Q^2_{ext} equal to 0.74). Kühne (19) developed a decision-tree model based on linear regression for the prediction of narcosis-level toxicity; read-across was then used to estimate the toxicity enhancement. Models were calibrated on 1,365 organic compounds and the final decision-tree provided a quantitative estimation for 757 compounds (56% of the data set) with a Q^2_{LOO} equal to 0.84.

Published models of acute toxicity toward *D. magna* have some drawbacks that can limit their actual application for regulatory purposes. One drawback, for instance, for PNNs and decision-tree models, is the complex modelling strategy. This can result in a difficult implementation, while OECD Principle 2 requires the "use of an unambiguous algorithm" in order to give transparency in the equations. Moreover, OECD Principle 5 requires a mechanistic interpretation, if possible: the model based on the PNN strategy lacks a direct mechanistic interpretation, due to both the large number of molecular descriptors (57 fragments) and the intrinsic complexity of the modelling algorithm. Also, OECD Principle 4, requires a correct validation procedure of the QSAR models, which, in some cases, is not properly fulfilled, since several published models were validated by optimistic procedures, such as the *ad hoc* selection of test molecules and leave-one-out cross validation (20).

In order to overcome the drawbacks and limitations of existing models, the aim of this study was to develop a QSAR model for the toxicity of organic chemicals toward *D. magna*, characterised by: a) a simple modelling method based on local structural

similarities; b) interpretable descriptors; c) an appropriate validation procedure to estimate the real predictivity and reliability of the model; and d) an implicit definition of the Applicability Domain (AD; 21, 22). In addition, attention was paid to data screening, in order to detect erroneous chemical structures and reduce the influence of anomalous toxicity values.

Materials and Methods

Experimental data

Experimental data on aquatic toxicity were retrieved from three databases (ECOTOX [23], EAT5 [24] and OASIS) and available scientific publications (25–41). The OASIS database was downloaded from the OECD QSAR Toolbox (42). The downloaded databases were imported into the Konstanz information miner (KNIME; 43), and *ad hoc*-designed workflows were used to extract LC50 data, which is the concentration that causes death in 50% of test *D. magna* over a test duration of 48 hours. Data were obtained under different experimental conditions, such as composition and characteristics (e.g. pH and temperature) of test water, test locations (laboratory or field), exposure types (e.g. static, flow-through, renewal). In the EAT5 database, LC50 data were reported as EC50 (effective concentration), with lethality as the observed effect. Records of the ECOTOX database indicating ranges or thresholds of experimental values were removed.

Data curation and filtering

In order to guarantee data consistency, data were checked, and ambiguous molecular structures and anomalous experimental values were disregarded.

Curation of molecular structures

Chemical names and CAS registry numbers (CASRNs) were available for every record in the data set. Web services to the chemical database, ChemSpider (44), and the Chemical Identifier Resolver (CIR; 45) of the CADD Group at NCI/NIH, were set up in the KNIME environment, to check the correctness of the molecular structures and the correspondence of CASRNs and names. CASRNs and chemical names were independently used as queries to retrieve the standard InChI codes and the Simplified Molecular-Input Line-Entry System (SMILES). The retrieved InChI codes were then compared. Out of 2,640 records (corresponding to 693 different CAS numbers), 1,577 (378 CAS numbers) presented mismatches. All the records that had at least one mismatch were manually checked by using the PubChem (46) and ChemSpider databases and the Sigma-Aldrich website (47). During this phase, some records were deleted for the following reasons: a) a chemical name–CASRN mismatch was not possible to resolve — for example, because the original publication was not found or was not accessible; b) the CASRN was non-existent; c) the molecular structure was not available, as it was a commercially-named chemical; d) information about which isomer was used was missing; and e) the record pertained to a chemical mixture. In total, 2,410 records, corresponding to 628 different CAS numbers were retained and merged with the data taken from scientific publications (195 records for a total of 183 different CAS numbers).

Filtering

The data set contained a certain number of disconnected structures, i.e. salts and mixtures. In particular, 733 records for a total of 118 disconnected structures were present. All the disconnected structures were removed from the data set, since toxic effects could arise from any of the chemical species present, either behaving independently or interacting to give additive, synergistic or antagonistic effects. Moreover, the calculation of molecular descriptors is limited when dealing with disconnected structures.

Inorganic compounds were removed, since the goal was to develop a model for acute toxicity that was limited to organic molecules. A total of 141 records, corresponding to 28 different inorganic compounds, were therefore removed.

Handling stereochemistry

Some stereoisomers were present in the data set. Since the majority of two-dimensional (2-D) molecular descriptors does not discriminate stereoisomers, the information about stereochemistry was removed from the SMILES before the calculation of molecular descriptors.

Curation of experimental values

Lethal concentrations were first converted to molarity and then transformed to a logarithmic scale (–Log mol/L). For several molecules, multiple values of LC50 were available, and in some cases, differences of a few orders of magnitude were observed for the same chemical. In order to avoid an excessive dependence on outlying data, the median value was calculated, as it is a more robust estimator than the mean value. The standard deviation was also calculated and used as an alert for

inconsistent data. The pooled standard deviation over the data set was equal to 0.37. Therefore, if the standard deviation of a molecule was larger than 0.7 log units (approximately twice the standard deviation over the entire data set), the original scientific publications were consulted in order to detect errors in the compilation of the databases. If the original study was not accessible or not found, the corresponding value was removed.

Experimental data for some Polycyclic Aromatic Hydrocarbons (PAHs) from a scientific publication (48) were removed, because toxicity had been photo-induced in the experimental tests.

The final data set included 546 organic molecules and is freely available (49, 50).

Molecular descriptors

The SMILES of the 546 organic molecules were used to calculate molecular descriptors. Three-dimensional (3-D) descriptors were not calculated, since the optimisation of molecular geometry may be a time-consuming step, and could also limit the future application of the model due to inconsistencies with the generation of 3-D structures (51).

One-dimensional and 2-D molecular descriptors implemented in the software DRAGON (52) were calculated. Constant, near-constant and descriptors with at least one missing value were removed, resulting in a total of 2,187 molecular descriptors.

Modelling methods

Due to the nature of the problem, non-linear regression methods were assumed to give better results than the classical linear regression. Methods based on local similarity are expected to be able to deal with non-linear responses, while still retaining a simple algorithm. This is the case for the *k*-Nearest Neighbour (*k*NN; 53) strategy, which was used to calibrate the models. The predicted value for a molecule is computed from the values of its *k* nearest neighbours, typically as a mean or weighted mean. In this study, the similarity between two molecules was calculated as:

$$S_{st} = \frac{1}{1 + d_{st}} = \frac{1}{1 + \sqrt{(\mathrm{x}_s - \mathrm{x}_t)^{\mathrm{T}}\ \mathrm{S}^{-1}\ (\mathrm{x}_s - \mathrm{x}_t)}} \quad 0 \le S_{st} \le 1$$

[Equation 1]

where d_{st} is the Mahalanobis distance between molecules *s* and *t*, x_s and x_t are the descriptor vectors for molecules *s* and *t*, and S^{-1} is the inverse of the covariance matrix of the training set. The predicted response, y_s, was computed as the weighted mean over the *k* neighbours, where the weights were calculated as a function of the similarity, as:

$$y_s = \sum_{t=1}^{k} y_t . w_t = \sum_{t=1}^{k} y_t . \frac{S_{st}}{\sum_{t=1}^{k} S_{st}}$$

[Equation 2]

where y_t and w_t are the response and the weight of the *t–th* neighbour, respectively, and the sum runs over the *k* neighbours. The term S_{st} is the similarity between molecules *s* and *t*, and the sum runs again over the *k* nearest neighbours.

A threshold value on the average distance from the *k* nearest neighbours was also adopted, in order to detect test molecules that are dissimilar from their *k* nearest neighbours. Hence, only molecules with an average distance from their neighbours lower than a defined threshold were predicted, while those exceeding the threshold were regarded as outliers on the assumption that their predictions may be influenced by dissimilar neighbours and therefore might not be reliable. The training molecules exceeding the threshold did not contribute to the model's statistics, but were not removed from the data set, since they still contributed to define the model's domain and, in principle, can be useful to predict the responses of external compounds.

Genetic Algorithms (GA) were coupled with *k*NN method to select the relevant molecular descriptors. The GA strategy described by Leardi and González was used in this study (54). For each combination of molecular descriptors (model), values of *k* (number of nearest neighbours) from 1 to 10 were tested. For each *k* value, the distance threshold from the *k* neighbours was automatically chosen during GA runs as the average distance value giving the largest coefficient of determination in cross-validation ($\mathrm{Q}^2_{\mathrm{cv}}$), with a constraint on the maximum allowed percentage of unpredicted molecules of 40%. This value was selected as a reasonable value to carry out the selection of molecular descriptors during model optimisation. Eventually, for each combination of molecular descriptors, the pair of *k* values and similarity threshold giving the largest $\mathrm{Q}^2_{\mathrm{cv}}$ was chosen as the optimal one.

Model validation

In order to thoroughly validate the developed models, the 546 molecules of the data set were randomly divided into a training set (436 molecules) and an external test set (110 molecules). The training set was used to calibrate models and select the optimal molecular descriptors by means of GA, while the test set was used only to test the predictive power of the calibrated models. During the GA runs, model performance was evaluated by means of internal five-fold cross-validation (55). The predictive ability on the external test set was evaluated by means of the $\mathrm{Q}^2_{\mathrm{ext}}$ function reported in the literature (56).

Software

KNIME (43) was used to process the databases, in order to extract the relevant data and check the molecular structures. Molecular descriptors were calculated by means of DRAGON 6 (52). Variable selection by means of GA, model fitting and validation were carried out in MATLAB (57), by using toolboxes and functions written by the authors.

Results and Discussion

The GA selection was organised into two subsequent steps, in order to handle the large number of calculated descriptors, i.e. 2,187, and to avoid potential over-fitting. Initially, GAs were run on each descriptor block separately. For each block, molecular descriptors with the largest frequencies of selection were chosen and merged together to form a set of 201 descriptors. GAs were then carried out on this reduced set, to find the most appropriate subset of descriptors.

Only one molecular descriptor, *TPSA(tot)* (topological polar surface area with N, O, S and P contributions; 58), had a selection frequency significantly larger than the others. In order to avoid selection based on small differences in the descriptor frequencies and to obtain a consistent solution, models based on the 15 most frequent descriptors were explored by means of an all-subset strategy, with two constraints: the maximum number of descriptors included in the models was set to 10; and *TPSA(tot)* was always included, since it proved to be relevant for the toxicity modelling. The best models were finally judged on the basis of both their predictive power and their complexity, also taking into consideration descriptor interpretability. This procedure resulted in a *k*NN model (*k* equal to three) constituted by eight molecular descriptors, which are briefly described below:

a) *MLOGP* is the octanol–water partition coefficient (LogP) calculated from the Moriguchi model (59, 60). LogP expresses the lipophilicity of a molecule, this being the driving force of narcosis.

b) *RDCHI* is a topological index (61) that encodes information about molecular size and branching, but does not account for heteroatoms. Since molecular size affects lipophilicity, it is reasonable that this descriptor also accounts, to a certain extent, for lipophilicity.

c) *SAacc* (62) describes the Van der Waals surface area (VSA) of atoms that are acceptors of hydrogen bonds.

d) *TPSA(tot)* (58) represents the topological polar surface area calculated by means of a contribution method that takes into account N, O, P and S. The two descriptors, *SAacc* and *TPSA(tot)*, taken together, account for the exposed molecular polar surface area that can interact with biological targets, where *SAacc* specifically takes into account the formation of hydrogen bonds, while the main contribution of *TPSA(tot)* is toward the calculation of the responses of P-containing and S-containing molecules (such as pesticides and herbicides).

e) *H-050* (63) represents the number of hydrogen atoms bonded to heteroatoms. Therefore, this descriptor still partly contains information related to the possibility of H-bond formation, but focuses on the number rather than on the surface area.

f) *nN* (4) is the number of nitrogen atoms present in the molecule. It is known that many nitrogen-containing functional groups are nucleophiles, due to the presence of a lone pair on the nitrogen atom (typically amines). Therefore, it is hypothesised that *nN* encodes information on the nucleophilicity, deriving from the presence of nitrogen atoms in the toxicants.

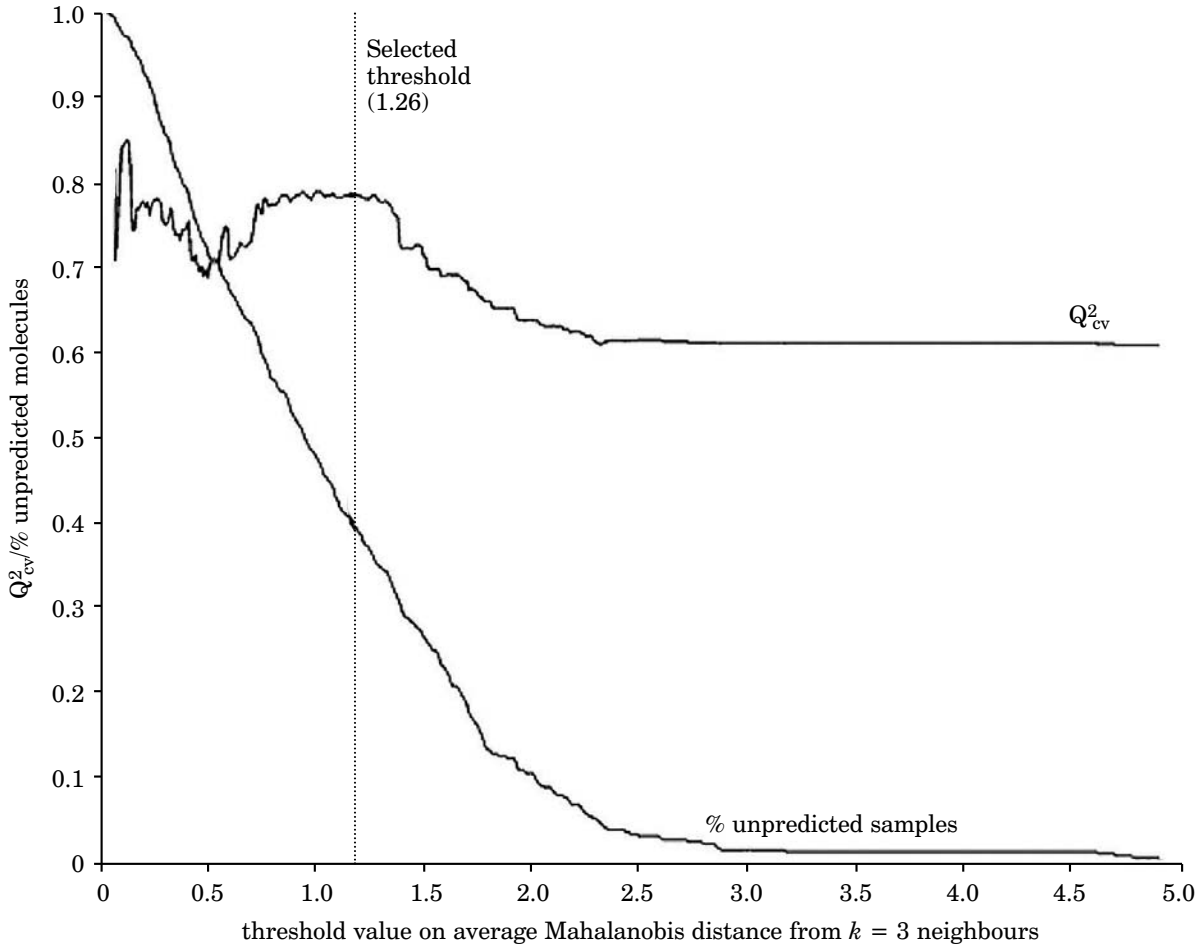
g) *C-040* (63) represents the number of carbon atoms of the type R–C(=X)–X / R–C#X / X=C=X, where X represents any electronegative atom (O, N, S, P, Se, halogens). In other words, *C-040* codifies specific functional groups such as esters, carboxylic acids, thioesters, carbamic acids, nitriles, etc. Since all of these groups are electron-poor on the carbon atoms, *C-040* seems to be able to account for electrophilic features.

h) *GATS1p* (4) encodes information on molecular polarisability, and tends to have low values for molecules with pairs of bonded atoms with comparable polarisabilities such as –OH, –NH and –NO. Moreover, *GATS1p* has smaller values when the polarisabilities of bonded atoms are large. In other words, the more polarisable a bond, the lower the value of *GATS1p*.

To conclude, the interpretation of the molecular descriptors was demonstrated to be in agreement with previous knowledge on the structural and electronic features that determine acute aquatic toxicity. It was confirmed that toxicity increases with lipophilicity, as a consequence of the enhanced ability of toxicants to enter the organism (narcosis). Moreover, a relationship was found between molecular polarisability and toxicity. This relationship was linked to the HSAB (64) and FMO (65) theories and the Klopman–Salem equation (66, 67), on the basis of the consideration that polarisable molecules are ‘soft’ species, which therefore tend to react with other soft species. In fact, it seems that more-polarisable molecules tend to have higher toxicities, and this might be due to the formation of covalent bonds that involve the HOMO and LUMO of soft acids and bases.

Figure 1 shows the Q^2_{cv} and percentage of unpredicted molecules as a function of the threshold. The percentage of unpredicted molecules decreased linearly with increasing threshold values, as was expected. On the other hand, model performance remained stable (Q^2_{cv} around 0.80) for threshold values in the range 0.8–1.4. A threshold value equal to 1.26 was finally selected as a reasonable trade-off between model predictivity and applicability limitation. Therefore, predictions for molecules with an

Figure 1: Q^2_{cv} and percentage of unpredicted samples as the function of the threshold value on the average Mahalanobis distance from $k = 3$ neighbours



The vertical line corresponds to the selected threshold value (1.26).

Table 2: Regression statistics of the *k*NN model

Model statistics							
k	**Av. dist threshold**	**R^2**	**Q^2_{cv}**	**Q^2_{ext}**	**% unpredicted fit**	**% unpredicted cv**	**% unpredicted test**
3	—	0.60	0.61	0.43	0	0	0
3	1.26	0.78	0.78	0.72	38	39	31

Q^2_{cv} = coefficient of determination in cross-validation; Q^2_{ext} = coefficient of determination in external validation; R^2 = coefficient of determination in fitting.

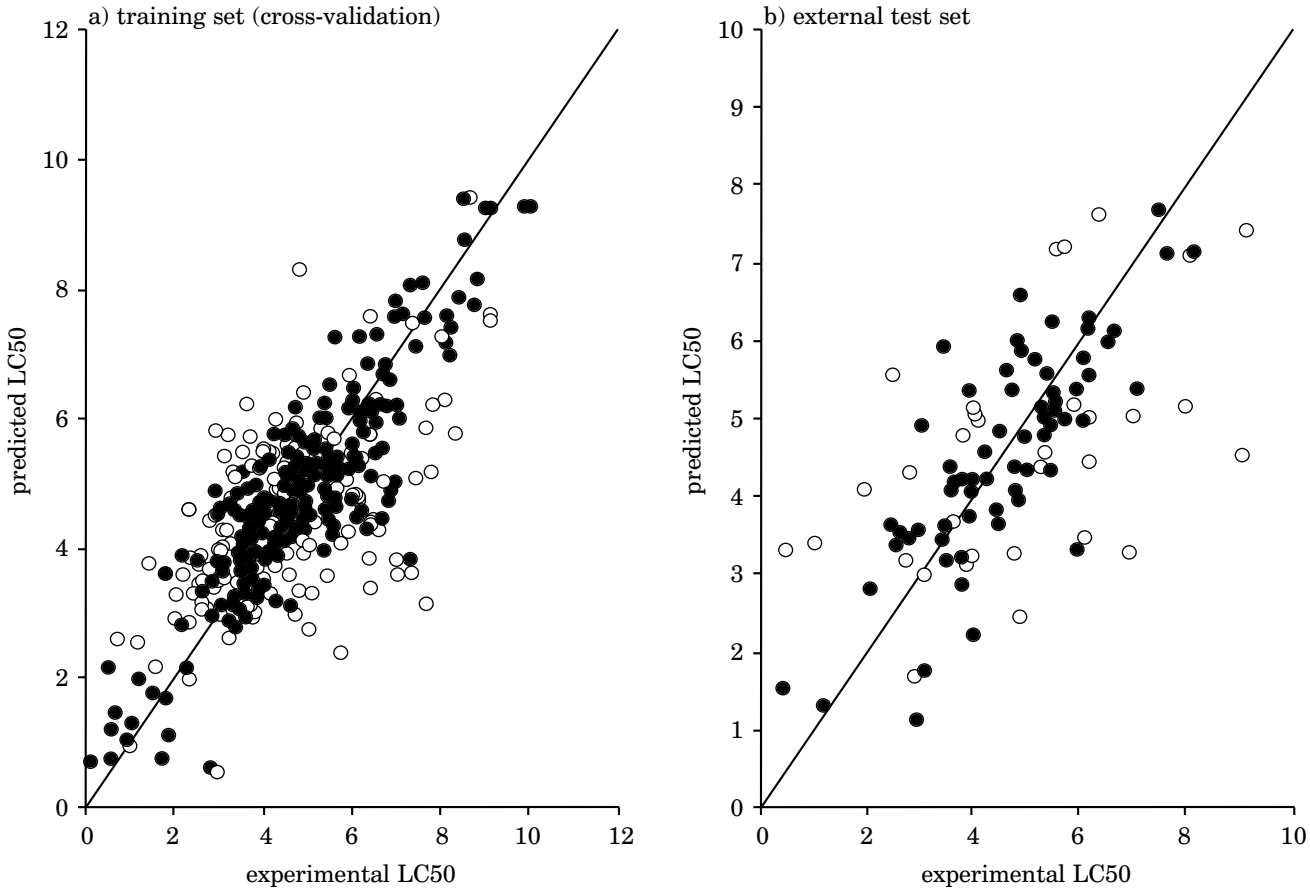
average distance from their three neighbours greater than 1.26 were regarded as unreliable and were not considered. If no threshold was considered, the classical *k*NN approach would be obtained with Q^2_{cv} equal to 0.61.

The threshold can also be user-defined to best suit the purposes of a specific study. For example, for high-throughput screening, where high reliability is not a strict requirement, one can increase the threshold value in order to have predictions for most of the molecules.

The developed QSAR model was finally validated on the external test set that was not part of the descriptor selection and model calibration. The regression statistics of the *k*NN model are collated in Table 2. The developed model was compared with a 'classical' *k*NN model where no molecule was left unpredicted. It is possible to see that the introduction of the threshold on the average distance enhanced the model's performance, since R^2, Q^2_{cv} and Q^2_{ext} were improved with 0.18, 0.17 and 0.29 points, respectively, to the detriment of the increase in the number of unpredicted molecules. Moreover, the performance of the model (as well as the percentage of unpredicted molecules) in fitting, cross-validation and external validation, gave similar values. This balance between model performance on the training and test sets indicates the absence of over-fitting, which can occur when dealing with variable selection on high-dimensional data.

Figure 2 shows the experimental *versus* predicted responses in cross-validation for the training set (Figure 2a) and for the external test set (Figure 2b). Black circles indicate compounds with average Mahalanobis distance from the three nearest neighbours which is lower than the selected threshold (i.e. 1.26). White circles indicate molecules with an average distance larger than the threshold. The introduction of the threshold per-

Figure 2: Experimental *versus* predicted responses for the training set and the external test set*



Black circles indicate compounds with average Mahalanobis distance from the three neighbours lower than the fixed threshold (1.26). White circles indicate molecules with average distance higher than the threshold.

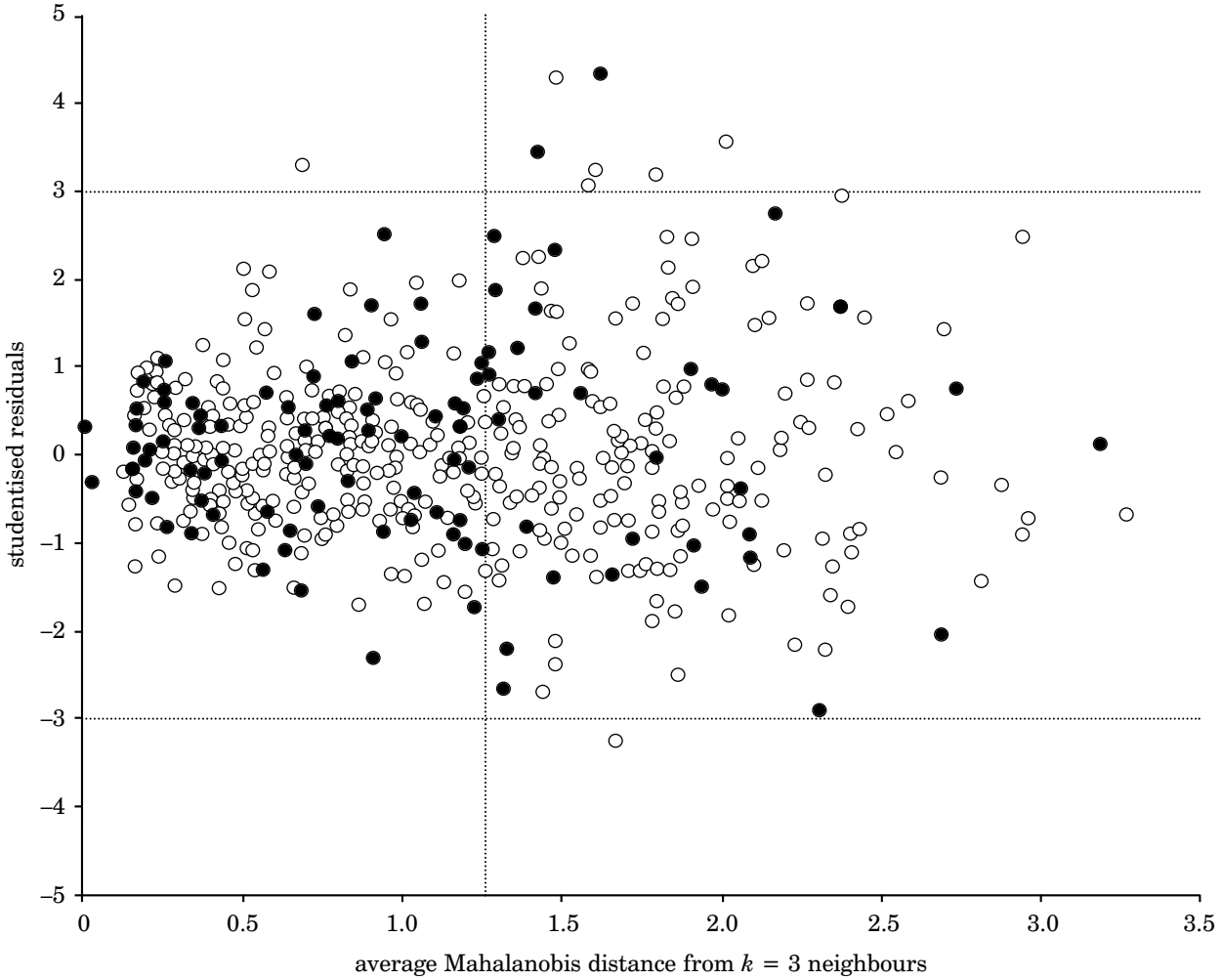
**This shows a corrected version of the graph in Figure 2a.*

mitted the identification of most of the molecules that were well predicted, while molecules very dissimilar from their neighbours also showed greater residuals in the response, especially in the case of the test set. Nevertheless, there are some molecules, with no structurally similar compounds, that are instead characterised by small residuals. This is likely to be the case of structural cliffs, i.e. molecules with different structures (small similarity) but similar toxicity.

Figure 3 shows, for each training and test molecule, the studentised residuals in cross-validation and external prediction *versus* the average distance from the three nearest neighbours. Six molecules had average Mahalanobis distances larger than 3.5, with a maximum value of 13.5. In order to make the plot more readable, the x-axis was cut at a value of 3.5. Predictions for molecules placed on the right hand-side of the vertical line (threshold value) were regarded as unreliable. A general trend of increasing residuals as the average distance increases can be observed.

Comparing the statistics of the proposed *k*NN model (Table 2) with those of other models calibrated on large heterogeneous data sets (Table 1), the proposed *k*NN model showed comparable performance, but was advantageous in the simplicity of its algorithm (OECD Principle 2), as well as its interpretability (OECD Principle 5). In fact, the proposed *k*NN model is based on only eight descriptors, while, for instance, the PNNs were based on 57 fragments. An additional important aspect for a QSAR model, especially when applied for regulatory purposes, is the definition of its AD,

Figure 3: Studentised residuals of the training set in cross-validation and the external test set *versus* the average Mahalanobis distance from the three neighbours



The vertical line represents the threshold value (1.26); the horizontal lines represent warning values on the residuals (3σ). ○ = *training set (cross-validation);* ● = *external test set.*

which is the chemical space where it can provide reliable predictions (OECD Principle 3). The introduction of a threshold on the average distance allowed the model to self-determine its own AD, because molecules with distances larger than the threshold are not predicted, on the assumption that their predictions are less (or not) reliable. Additional advantages that the *k*NN model can provide are that it allows a local similarity analysis based on the nearest neighbours for each molecule to be predicted, and it can include new molecules in the training set without the need for recalculation of model parameters, except for the covariance matrix.

Conclusions

This study addressed the problem of predicting the toxicity of organic chemicals toward *D. magna* by means of a QSAR model that was developed to comply with the OECD principles required for the model to be applicable for regulatory purposes.

Data on aquatic toxicity (LC50 on *D. magna* over a test duration of 48 hours) were taken from three databases and 17 additional scientific publications (25–41). *Ad hoc*-designed workflows were used for data curation and filtering. The final data set comprised 546 organic molecules, randomly divided into a training set and an external test set. The GA-*k*NN strategy was implemented with a threshold on the average Mahalanobis distance from the *K* nearest neighbours, so that only molecules satisfying the threshold criterion were predicted. The final QSAR model showed good performance in fitting (R^2 equal to 0.78), cross-validation (Q^2_{cv} equal to 0.78) and external prediction (Q^2_{ext} equal to 0.72), with percentages of unpredicted molecules equal to 38%, 39%, and 31% in fitting, cross-validation and external validation, respectively. An analysis of the residuals on both the training and test sets showed that high residuals were associated with large average distances from the neighbours, thus justifying the introduction of the threshold. The model comprised eight molecular descriptors that encoded information about lipophilicity, formation of H-bonds, polar surface area, polarisability, nucleophilicity and electrophilicity.

Acknowledgment

The research leading to these results has partly been financed by the EU Seventh Framework Programme Marie Curie Initial Training Network Environmental ChemOinformatics (ECO; under Grant Agreement No. 238701).

References

1. Newsome, L.D., Nabholz, J.V. & Kim, A. (1996). Designing aquatically safer chemicals. In *Designing Safer Chemicals: Green Chemistry for Pollution Prevention* (ed. S.C. DeVito & R.L. Garrett), pp. 172–192. Washington, DC, USA: American Chemical Society.
2. Rand, G.M. & Petrocelli S.R. (1985). *Fundamentals of Aquatic Toxicology: Methods and Applications*, 666pp. Washington, DC, USA: Hemisphere Publishing.
3. European Parliament (2006). *Regulation (EC) No 1907/2006* of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending *Directive 1999/45/EC* and repealing *Council Regulation (EEC) No 793/93* and *Commission Regulation (EC) No 1488/94* as well as *Council Directive 76/769/EEC* and *Commission Directives 91/155/EEC*, *93/67/EEC*, *93/105/EC* and *2000/21/EC*. *Official Journal of the European Union* **L396**, 30.12.2006, 1–849.
4. Todeschini, R. & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*, 1257pp. Weinheim, Germany: Wiley-VCH.
5. OECD (undated). *The Organisation for Economic Co-operation and Development (OECD)*. [Homepage.] Available at: http://www.oecd.org (Accessed 10.01.14).
6. Vighi, M., Masoero Garlanda, M. & Calamari, D. (1991). QSARs for toxicity of organophosphorous pesticides to *Daphnia* and honeybees. *Science of the Total Environment* **109/110**, 605–622.
7. Vighi, M. & Calamari, D. (1985). QSARs for organotin compounds on *Daphnia magna*. *Chemosphere* **14**, 1925–1932.
8. Todeschini, R., Vighi, M., Provenzani, R., Finizio, A. & Gramatica, P. (1996). Modeling and prediction by using WHIM descriptors in QSAR studies: Toxicity of heterogeneous chemicals on *Daphnia magna*. *Chemosphere* **32**, 1527–1545.
9. Deneer, L.W., van Leeuwen, C.J., Seinen, W., Maas-Diepveen, J.L. & Hermens, J.L.M. (1989). QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna*, *Chlorella pyrenoidosa* and *Photobacterium phosphoreum*. *Aquatic Toxicology* **15**, 83–98.
10. Hossain, M.I., Samir, B.B., El-Harbawi, M., Masri, A.N., Mutalib, M.I.A., Hefter, G. & Yin, C.Y. (2011). Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities. *Chemosphere* **85**, 990–994.
11. Zvinavashe, E., Du, T., Griff, T., van den Berg, H.H.J., Soffers, A.E.M.F., Vervoort, J., Murk, A.J. & Rietjens, I.M.C.M. (2009). Quantitative structure–activity relationship modeling of the toxicity of organothiophosphate pesticides to *Daphnia magna* and *Cyprinus carpio*. *Chemosphere* **75**, 1531–1538.
12. Cassani, S., Kovarich, S., Papa, E., Roy, P.P., van der Wal, L. & Gramatica, P. (2013). *Daphnia* and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling. *Journal of Hazardous Materials* **258/259**, 50–60.
13. Cassani, S., Kovarich, S., Papa, E., Roy, P.P., Rahmberg, M., Nilsson, S., Sahlin, U., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I.V., Brandmaier, S.,

Durjava, M.K., Kolar, B., Peijnenburg, W. & Gramatica, P. (2013). Evaluation of CADASTER QSAR models for aquatic toxicity of (benzo)triazoles and prioritisation by consensus. *ATLA* **41**, 49–64.
14. Tetko, I.V., Sopasakis, P., Kunwar, P., Brandmaier, S., Novotarskyi, S., Charochkina, L., Prokopenko, C. & Peijnenburg, W.J.G.M. (2013). Prioritization of polybrominated diphenyl ethers (PBDEs) using the QSPR-THESAURUS web tool. *ATLA* **41**, 127–135.
15. Faucon, J.C., Bureau, R., Faisant, J., Briens, F. & Rault, S. (2001). Prediction of the *Daphnia* acute toxicity from heterogeneous data. *Chemosphere* **44**, 407–422.
16. Katritzky, A.R., Slavov, S.H., Stoyanova-Slavova, I.S., Kahn, I. & Karelson, M. (2009). Quantitative structure–activity relationship (QSAR) modeling of EC50 of aquatic toxicities for *Daphnia magna*. *Journal of Toxicology & Environmental Health, Part A* **72**, 1181–1190.
17. Kaiser, K.L.E. & Niculescu, S.P. (2001). Modeling acute toxicity of chemicals to *Daphnia magna*: A probabilistic neural network approach. *Environmental Toxicology & Chemistry* **20**, 420–431.
18. Kar, S. & Roy, K. (2010). QSAR modeling of toxicity of diverse organic chemicals to *Daphnia magna* using 2D and 3D descriptors. *Journal of Hazardous Materials* **177**, 344–351.
19. Kühne, R., Ebert, R.U., von der Ohe, P.C., Ulrich, N., Brack, W. & Schüürmann, G. (2013). Read-across prediction of the acute toxicity of organic compounds toward the water flea *Daphnia magna*. *Molecular Informatics* **32**, 108–120.
20. Golbraikh, A. & Tropsha, A. (2002). Beware of Q2! *Journal of Molecular Graphics & Modelling* **20**, 269–276.
21. Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V. & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **17**, 4791–4810.
22. Tetko, I.V., Bruneau, P., Mewes, H.W., Rohrer, D.C. & Poda, G.I. (2006). Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* **11**, 700–707.
23. US EPA (2013). *ECOTOX Database, Version 4.0*. Washington, DC, USA: US Environmental Protection Agency. Available at: http:/www.epa.gov/ecotox/ (Accessed 20.12.13)
24. ECETOC (2003). *Technical Report 091*: *Aquatic Hazard Assessment II*, pp. 1–164. Brussels, Belgium: European Centre for Ecotoxicology and Toxicology of Chemicals. Available at: http://www.ecetoc.org/technical-reports (04.01.14).
25. Bernot, R.J., Brueseke, M.A., Evans-White, M.A. & Lamberti, G.A. (2005). Acute and chronic toxicity of imidazolium-based ionic liquids on *Daphnia magna*. *Environmental Toxicology & Chemistry* **24**, 87–92.
26. Randall, W.F., Dennis, W.H. & Warner, M.C. (1979). Acute toxicity of dechlorinated DDT, chlordane and lindane to bluegill (*Lepomis macrochirus*) and *Daphnia magna*. *Bulletin of Environmental Contamination & Toxicology* **21**, 849–854.
27. Sanderson, H. & Thomsen, M. (2009). Comparative analysis of pharmaceuticals *versus* industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action. *Toxicology Letters* **187**, 84–93.
28. Jemec, A., Tisler, T., Drobne, D., Sepcic, K., Fournier, D. & Trebse, P. (2007). Comparative toxicity of imidacloprid, of its commercial liquid formulation and of diazinon to a non-target arthropod, the microcrustacean *Daphnia magna*. *Chemosphere* **68**, 1408–1418.
29. Zou, E. & Fingerman, M. (1997). Effects of estrogenic xenobiotics on molting of the water flea, *Daphnia magna*. *Ecotoxicology & Environmental Safety* **38**, 281–285.
30. Costanzo, S.D., Watkinson, A.J., Murby, E.J., Kolpin, D.W. & Sandstrom, M.W. (2007). Is there a risk associated with the insect repellent DEET (*N*,*N*-diethyl-*m*-toluamide) commonly found in aquatic environments? *Science of the Total Environment* **384**, 214–220.
31. Staples, C.A. & Davis, J.W. (2002). An examination of the physical properties, fate, ecotoxicity and potential environmental risks for a series of propylene glycol ethers. *Chemosphere* **49**, 61–73.
32. Martins, J.C., Saker, M.L., Oliva Teles, L.F. & Vasconcelos, V.M. (2007). Oxygen consumption by *Daphnia magna* Straus as a marker of chemical stress in the aquatic environment. *Environmental Toxicology & Chemistry* **26**, 1987–1991.
33. Von der Ohe, P.C., Kuhne, R., Ebert, R., Altenburger, R., Liess, M. & Schuurmann, G. (2005). Structural alerts — A new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. *Chemical Research in Toxicology* **18**, 536–555.
34. Williams, E.S., Berninger, J.P. & Brooks, B. (2011). Application of chemical toxicity distributions to ecotoxicology data requirements under REACH. *Environmental Toxicology & Chemistry* **30**, 1943–1954.
35. Nørgaard, K.B. & Cedergreen, N. (2010). Pesticide cocktails can interact synergistically on aquatic crustaceans. *Environmental Science & Pollution Research* **17**, 957–967.
36. Dojmi di Delupis, G., Macri, A., Civitareale, C. & Migliore, L. (1992). Antibiotics of zootechnical use: Effects of acute high and low dose contamination on *Daphnia magna* Straus. *Aquatic Toxicology* **22**, 53–60.
37. Ferrari, B., Mons, R., Vollat, B., Fraysse, B., Paxeus, N., Lo Giudice, R., Pollio, A. & Garric, J. (2004). Environmental risk assessment of six human pharmaceuticals: Are the current environmental risk assessment procedures sufficient for the protection of the aquatic environment? *Environmental Toxicology & Chemistry* **23**, 1344–1354.
38. Foit, K., Kaske, O. & Liess, M. (2012). Competition increases toxicant sensitivity and delays the recovery of two interacting populations. *Aquatic Toxicology* **106/107**, 25–31.
39. Ochoa-Acuna, H.G., Bialkowski, W., Yale, G. & Hahn, L. (2009). Toxicity of soybean rust fungicides to freshwater algae and *Daphnia magna*. *Ecotoxicology* **18**, 440–446.
40. Horn, O., Nalli, S., Cooper, D. & Nicell, J. (2004). Plasticizer metabolites in the environment. *Water Research* **38**, 3693–3698.
41. Kyriakopoulou, K., Anastasiadou, P. & Machera, K. (2009). Comparative toxicities of fungicide and herbicide formulations on freshwater and marine species. *Bulletin of Environmental Contamination &*

Toxicology **82**, 290–295.
42. OECD (2012). *The OECD QSAR Toolbox for Grouping Chemicals into Categories, Version 2.3*. Paris, France: Organisation for Economic Co-operation and Development. Available at: http://www.qsartoolbox.org/ (Accessed 04.12.13).
43. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K. & Wiswedel, B. (2007). KNIME: The Konstanz information miner. In *Studies in Classification, Data Analysis and Knowledge Organization*, pp. 319–326. London, UK: Springer. [ISSN: 1431-8814.]
44. RSC (2013). *ChemSpider*. Cambridge, UK: Royal Society of Chemistry. Available at: http://www.chemspider.com/ (Accessed 20.12.13).
45. NCI/CADD Group (2013). *Chemical Identifier Resolver*. Available at: http://cactus.nci.nih.gov/chemical/structure (Accessed 20.12.13).
46. Bolton, E., Wang, Y., Thiessen, P.A. & Bryant, S.H. (2008). Chapter 12 PubChem: Integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry* **4**, 217–241.
47. Anon. (2012). *Sigma-Aldrich Co.* [Homepage.] Available at: http://www.sigmaaldrich.com (Accessed 20.12.13).
48. Lampi, M.A., Gurska, J., McDonald, K.I.C., Xie, F., Huang, X.D., Dixon, D.G. & Greenberg, B.M. (2005). Photoinduced toxicity of polycyclic aromatic hydrocarbons to *Daphnia magna*: Ultraviolet-mediated effects and the toxicity of polycyclic aromatic hydrocarbon photoproducts. *Environmental Toxicology & Chemistry* **25**, 1079–1087.
49. Todeschini, R. (undated). *Acute Aquatic Toxicity Dataset*. Milan, Italy: Milano Chemometrics and QSAR Research Group. Available at: http://michem.disat.unimib.it/chm/download/toxicity.htm (Accessed 20.12.13).
50. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V.V., Tanchuk, V.Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I.I., Palyulin, V.A., Radchenko, E.V., Welsh, W.J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-de-Sousa, J., Zhang, Q.Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V. & Tetko, I.V. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *Journal of Computer-Aided Molecular Design* **25**, 533–554.
51. Brandmaier, S., Peijnenburg, W., Durjava, M.K., Kolar, B., Gramatica, P., Papa, E., Bhhatarai, B., Kovarich, S., Cassani, S., Roy, P.P., Rahmberg, M., Öberg, T., Jeliazkova, N., Golsteijn, L., Comber, M., Charochkina, L., Novotarskyi, S., Sushko, I., Abdelaziz, A., D'Onofrio, E., Kunwar, P., Ruggiu, F. & Tetko, I.V. (2014). The QSPR-THESAURUS: The online platform of the CADASTER project. *ATLA* **42**, 13–24.
52. Talete srl (2013). *Talete srl, Dragon (Software for Molecular Descriptor Calculation) Version 6.0 — 2013*. Available at: http://www.talete.mi.it/ (Accessed 12.03.14).
53. Kowalski, B.R. & Bender, C.F. (1972). The K-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Analytical Chemistry* **44**, 1405–1411.
54. Leardi, R. & González, A.L. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemometrics & Intelligent Laboratory Systems* **41**, 195–207.
55. Cruciani, G., Baroni, M., Costantino, G., Riganelli, D. & Skagerberg, B. (1992). Predictive ability of regression models. Part I: Standard deviation of prediction errors (SDEP). *Journal of Chemometrics* **6**, 335–346.
56. Consonni, V., Ballabio, D. & Todeschini, R. (2009). Comments on the definition of the Q^2 parameter for QSAR validation. *Journal of Chemical Information & Modeling* **49**, 1669–1678.
57. Anon. (undated). *MATLAB R2012a (64-bit)*. Natick, MA, USA: MathWorks Inc.
58. Ertl, P., Rohde, B. & Selzer P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry* **43**, 3714–3717.
59. Moriguchi, I., Hirono, S., Nakagome, I. & Hirano, H. (1994). Comparison of reliability of log P values for drugs calculated by several methods. *Chemical & Pharmaceutical Bullettin* **42**, 976–978.
60. Moriguchi, I., Hirono, S., Liu, Q., Nakagome, I. & Matsushita, Y. (1992). Simple method of calculating octanol/water partition coefficient. *Chemical & Pharmaceutical Bullettin* **40**, 127–130.
61. Ivanciuc, O., Balaban, T.S. & Balaban, A.T. (1993). Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. *Journal of Mathematical Chemistry* **12**, 309–318.
62. Labute, P. (2000). A widely applicable set of descriptors. *Journal of Molecular Graphics & Modelling* **18**, 464–477.
63. Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. & Robins, R.K. (1989). Atomic physicochemical parameters for three dimensional structure directed quantitative structure–activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of Chemical Information & Modelling* **29**, 163–172.
64. Pearson, R.G. (1963). Hard and soft acids and bases. *Journal of the American Chemical Society* **85**, 3533–3539.
65. Fukui, K., Yonezawa, T. & Shingu, H. (1952). A molecular orbital theory of reactivity in aromatic hydrocarbons. *Journal of Chemical Physics* **20**, 722.
66. Klopman, G. (1968). Chemical reactivity and the concept of charge- and frontier-controlled reactions. *Journal of the American Chemical Society* **90**, 223–234.
67. Salem, L. (1968). Intermolecular orbital theory of the interaction between conjugated systems. I. General theory. *Journal of the American Chemical Society* **90**, 543–552.